VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

HLEDÁNÍ VHODNÝCH DESKRIPTORŮ PRO POPIS BAKTERIÁLNÍCH GENOMŮ

SEARCH OF THE BACTERIAL GENOME DESCRIPTOR

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **ROMAN VANEK**
AUTHOR

VEDOUCÍ PRÁCE **Ing. IVANA BURGETOVÁ, Ph.D.**
SUPERVISOR

BRNO 2017

Zadání bakalářské práce/19659/2016/xvanek36

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav informačních systémů Akademický rok 2016/2017

Zadání bakalářské práce

Řešitel: **Vanek Roman**
Obor: Informační technologie
Téma: **Hledání vhodných deskriptorů pro popis bakteriálních genomů**
Search of the Bacterial Genome Descriptor
Kategorie: Bioinformatika

Pokyny:

1. Seznamte se s databázemi DNA sekvencí a s databázemi genomů.
2. Po dohodě s vedoucí připravte DNA sekvence různých druhů bakterií.
3. Implementujte nástroj, který umožní analýzu složení zadaných DNA sekvencí (analýzu počtu různých n-tic nukleotidů) a výsledky zobrazí vhodným způsobem.
4. Zjistěte, zda existují nějaké rozdíly ve složení připravených sekvencí.
5. Zhodnoťte dosažené výsledky.

Literatura:

- PEVSNER, Jonathan. *Bioinformatics and functional genomics*. 2nd ed. Hoboken, N.J: Wiley-Blackwell, 2009, ISBN 978-0-470-08585-1.
- GENTLEMAN, Robert. *Bioinformatics and computational biology solutions using R and bioconductor*. New York: Spinger Science Business Media, 2005, ISBN 0-387-25146-4.

Pro udělení zápočtu za první semestr je požadováno:

- Body 1 a 2.

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Burgetová Ivana, Ing., Ph.D.**, UIFS FIT VUT
Datum zadání: 1. listopadu 2016
Datum odevzdání: 17. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav informačních systémů
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Dušan Kolář
vedoucí ústavu

Abstrakt

Táto bakalárska práca sa zaoberá skúmaním DNA sekvencií genómov rôznych baktérií. Cieľom je určiť, či je možné na základe počtov n-tíc nukleotidov zistiť, z ktorej baktérie, resp. druhu baktérií sekvencia pochádza. Výsledky skúmania boli získané pomocou nástroja, ktorého realizácia tvorí súčasť práce. Výstupom tejto práce sú experimenty so sekvenciami na báze štatistických výpočtov s frekvenciami n-tíc a rozhodnutie, či je tento prístup prínosný.

Abstract

This bachelor thesis deals with the investigation of DNA sequences of bacterial genomes. The aim is to find out whether it is possible to determine on the basis of the nucleotide tuple frequencies from which bacteria specie the examined sequence originates. The investigation results were obtained by a tool whose implementation is also a part of the work. The output of this thesis are the experiments with nucleotide tuple frequencies based on statistical calculations and decision whether this approach is beneficial or not.

Kľúčové slová

bakteriálny genóm, biologická databáza, nukleotid, fylogenetický strom

Keywords

bacterial genome, biological database, nucleotide, phylogenetic tree

Citácia

VANEK, Roman. *Hledání vhodných deskriptorů pro popis bakteriálních genomů*. Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Ivana Burgetová, Ph.D.

Hledání vhodných deskriptorů pro popis bakteriálních genomů

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pani Ing. Ivany Burgetovej, Ph.D. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Roman Vanek
17. mája 2017

Poďakovanie

Rád by som sa touto cestou poďakoval vedúcej práce Ing. Ivane Burgetovej, Ph.D. za odbornú pomoc, prínosné nápady a trpezlivosť počas realizácie tejto práce.

Obsah

Kapitola 1

Úvod

Táto práca sa zaoberá analýzou sekvencií bakteriálnych genómov. Dĺžky týchto sekvencií sa pohybujú v rádoch miliónov nukleotidov a ich porovnávanie býva časovo i výpočtovo náročné. Cieľom práce je pokúsiť sa nájsť zjednodušený popis genómov založený na rôznych počtoch n-tíc nukleotidov a zistiť, či frekvencie nukleotidov postačia k tomu, aby sme mohli vyvodiť užitočné závery, napr. rozhodnúť, z ktorej skupiny baktérií konkrétna baktéria pochádza.

Taktiež budú skúmané možnosti popisu náhodných DNA sekvencií, ku ktorým nie sú známe žiadne dodatočné informácie. Z týchto sekvencií prirodzene nemôžeme určiť žiadne parametre okrem ich dĺžky a frekvencií rôznych n-tíc nukleotidov.

Práca je určená najmä pre tých, ktorí pracujú s DNA sekvenciami a hľadajú rôzne možnosti zjednodušenia prístupu, ale aj pre každého zvedavca, ktorého táto tematika zaujíma.

Súčasťou práce je nástroj, ktorý umožňuje analýzu daných sekvencií fungujúci na báze skriptov. Nástroj dokáže spracovať ľubovoľné sekvencie nukleotidov zadané vo formátoch FASTA s prípadnou anotáciou vo formáte GFF3 a takisto na základe frekvencií nukleotidov generuje štruktúru fylogenetického stromu.

V kapitole 2 je sú na úvod objasnené základy molekulárnej biológie a stručný prehľad o baktériách a ich štruktúre. V kapitole 3 je predstavený úvod do biologických databáz, sú popísané možnosti získania biologických dát a príklady práce s nimi. Nasledujúca kapitola popisuje konkrétny návrh nástroja využitého v tejto práci. Posledná kapitola obsahuje experimenty a ich zhodnotenie.

Kapitola 2

Základné pojmy molekulárnej biológie

Na úvod do problematiky budú v tejto kapitole objasnené základné pojmy molekulárnej biológie ako je bunka, molekula DNA, gén, genóm a následne základný prehľad o baktériách a všeobecnom rozdeľovaní organizmov na báze vývojovej línie.

2.1 Molekula DNA a jej štruktúra

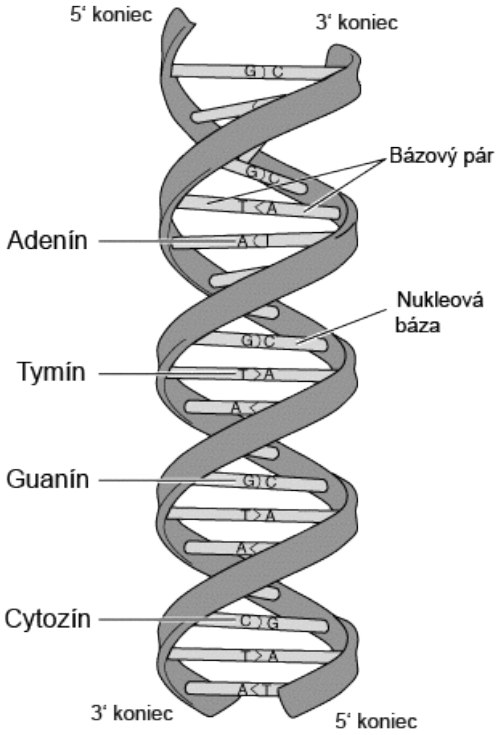
Základným stavebným prvkom prevažnej časti živých organizmov je bunka. Bunka svojou individualitou pripomína obyvateľa krajiny, kde v tomto prípade krajinu pripomína určitý orgán alebo celý organizmus. Samotná je prehliadnuteľná, je len jedna z mnohých, no spojením s inými bunkami tvoria súdržný celok. Bunka slúži ako identifikátor organizmu, tj. dokážeme podľa ľubovoľnej bunky určiť, z akého druhu organizmu pochádza. Hovorí sa, že „časť celku je časťou preto, lebo vo svojej individualite vykazuje charakteristiky celku“ [14].

Bunka obsahuje genetickú informáciu o organizme. Tú reprezentuje molekula DNA (deoxyribonukleová kyselina; angl. *deoxyribonucleic acid*; na obr. 2.1). Podľa typu umiestnenia molekuly DNA v bunke delíme organizmy na prokaryoty a eukaryoty. Prokaryoty sú staršie a jednoduchšie organizmy a molekulu DNA majú uloženú priamo v cytoplazme bunky. Eukaryotické organizmy sú štrukturálne zložitejšie a v ich bunke sa nachádza jadro. Jadro vykonáva funkciu „koordinačného centra“, čiže to funguje podobne ako v pracovnom procese, kde sa kontrola nad väčším množstvom zamestnancov rieši nadriadeným, ktorý zastrešuje menší celok. V tomto jadre sa nachádza genetická informácia.

Molekula DNA je polymér[1] a tvorí ho reťazec nukleotidov. Nukleotid pripomína tehlu v konštrukcii budovy. Tvoria ho 3 zložky plniace rozličné funkcie: sacharidová zložka, fosfátová zložka a nukleová báza. Sacharidová a fosfátová zložka hrajú úlohu napr. pri prenose energie v bunke. Nukleové bázy sú 4, menovite adenín (skratka A), cytozín (C), guanín (G) a tymín (T). U rastlín sa namiesto tymínu nachádza uracil (U). Nukleotidy tvoria dlhé, tzv. polynukleotidové reťazce. Trojice nukleotidov so zvláštnou funkciou sa nazývajú kodóny.

Molekulu DNA tvoria dva polynukleotidové reťazce. Tie sú usporiadané do pravotočivej dvojzávitnice (angl. *double helix*). Reťazce sú pevne spojené a sú si voči sebe komplementárne. Komplementárne (doplnkové) bázy sú kombinácie A – T a C – G (pozri obr. 2.2), takže komplementárny reťazec k reťazcu AACTG bude TTGAC. Takéto „zdvojenie“ pripomína v informatike rôzne zabezpečenia proti chybám, čiže podobne ako kontrolný súčet umožňuje

[1] makromolekula, dlhý reťazec



Obr. 2.1: Molekula DNA [25]

pri chýbajúcom údaji pôvodnú informáciu zrekonštruovať. Adenín a guanín majú purínový[2] základ a cytozín s tymínom pyrimidínový základ. Purín sa dobre viaže s pyrimidínom cez vodíkovú väzbu (tiež vodíkový mostík) a tým je realizovaná komplementárnosť.

Štruktúra dvojzávitnice bola popísaná v roku 1953 Jamesom D. Watsonom a Francisom Crickom, ktorí za tento objav dostali roku 1962 Nobelovu cenu [16]. Špirálovité útvary sa vo vesmíre nachádzajú v hojnom počte. Napríklad galaxie majú väčšinou špirálovitý tvar. Taktiež aj pohyb Zeme vzhľadom k slnečnej sústave je špirálovitého charakteru. Výhodou špirálovitej štruktúry je veľká odolnosť a trvanlivosť. K prerušeniu vodíkových mostíkov u nukleotidov však stačí zvýšená teplota (90 až 100 °C).

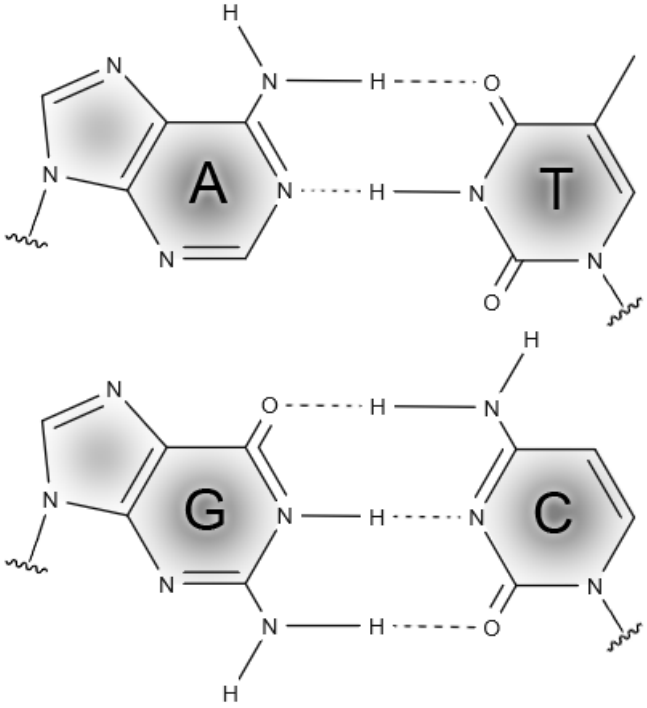
Reťazec molekuly má prirodzene dva konce, ktoré značíme ako 5’ koniec a 3’ koniec. Tieto označenia súvisia s miestom naviazania fosfátovej zložky nukleotidu a pomáhajú pri určovaní smeru molekuly. Dva reťazce molekuly sú teda voči sebe komplementárne a antiparalelné, tzn. 5’ koniec jedného reťazca sa viaže s 3’ koncom druhého a naopak.

Pre popis dĺžky molekuly sa využíva jednotka bázový pár (angl. *base pair*) so skratkou **bp**, ktorý značí pár dvoch komplementárnych báz. Z hľadiska dĺžky molekuly má tento *pár* dĺžku rovnú 1, keďže do úvahy sa berie len jeden prvok z dvojice. Násobné jednotky sú:

$$\begin{aligned} 1\,\mathrm{kbp} &= 10^3\,\mathrm{bp} \\ 1\,\mathrm{Mbp} &= 10^3\,\mathrm{kbp} = 10^6\,\mathrm{bp}, \end{aligned}$$

nazývané *kilobáza* a *megabáza*, ktoré sú často využívané kvôli veľkým dĺžkam sekvencií – v rádoch miliónov bázových párov.

[2] Purín a pyrimidín sú heterocyklické zlúčeniny obsahujúce atómy dusíka (načrtnuté na obr. 2.2)



Obr. 2.2: Komplementárnosť nukleových báz využívajúca vodíkové väzby (znázornené čiarkovane). „Vlnka“ označuje miesto naviazania sacharidovej zložky.

Prokaryotické organizmy (baktérie) sú jednobunkové a ich DNA je uložená v cytoplazme, kde vytvára kruhový útvar. U eukaryotických buniek DNA tvorí zväčša lineárne útvary. Útvary z molekuly DNA sa nazývajú chromozómy. Chromozóm môže byť v jednej bunke buď jeden, alebo viacero[3]. Prokaryotá však obvykle obsahujú jeden chromozóm, ktorý býva označovaný aj pod pojmom nukleoid.

Prokaryoty tiež môžu obsahovať malé kruhovité molekuly – plazmidy. Plazmidy majú veľkosť približne 1 000 až 200 000 báz [21]. Plazmidové gény riadia niektoré špecifické vlastnosti bakteriálnych buniek, napr. odolnosť voči antibiotikám [8].

2.2 Gén a genóm ako serializovaný organizmus

Každý človek sa vo svojom živote bežne stretáva s pojmom gén, avšak mnohým je podstata tohoto termínu neznáma. Je ale všeobecne známe, že ak sa u človeka prejavuje nejaké nadanie alebo iná zvláštnosť, reakcia okolia je zväčša: „má to v génoch“. Z tejto zaužívanej frázy možno predpokladať, že gén predstavuje určitú štruktúru vyskytujúcu sa v organizme, ktorá určuje nejakú jeho vlastnosť.

Slovo gén sa objavilo roku 1909, keď ho dánsky bádateľ Wilhelm Johansen použil k označeniu materiálneho nosiča dedičnosti [8]. Je možné toto slovo chápať dvojako. Za prvé, gén je základnou informačnou jednotkou dedičnosti. T. H. Morgan pri štúdiu vínnych mušiek (*Drosophila melanogaster*) dospel k záveru, že chromozómy sú nositele genetickej informácie, tj. gén úzko súvisí s určitým úsekom chromozómu [26]. Keďže molekula DNA je

[3]Človek má 23 párov, dokopy 46 chromozómov.

usporiadaná v chromozómoch, tak druhý význam slova gén možno chápať tak, že gén je záznam genetickej informácie v konkrétnom úseku molekuly DNA.

Gény (ako úseky DNA) sú významovo ucelené jednotky, ktoré hrajú úlohu pri tvorbe proteínov [8]. Proteíny už určujú vlastnosti, ktoré daný organizmus bude mať, napr. farba kvetov u rastlín. Tvorba proteínov z DNA sa nazýva proteosyntéza.

Genóm je celkový súbor genetickej informácie o organizme. Je to teda súbor všetkých génov a ostatných častí DNA, ktorý presne identifikuje organizmus. Na tomto princípe funguje tzv. test DNA, keď sa podľa vzorky (napr. vlasu) určuje jej (už bývalý) „vlastník“. Rozsah genómu je a priori priamo úmerný zložitosti štruktúry organizmu. Genóm najmenších vírusov neobsahuje ani 10 génov, no na druhej strane, genóm niektorých eukaryotických organizmov je rozsiahly. Genóm človeka obsahuje približne 20 tisíc génov [16], no tie sú v celej DNA rozložené veľmi riedko. Väčšinu sekvencie tvoria úseky zvané intróny, ktoré sa ďalej netransformujú do proteínov a ich význam nie je do dnešného dňa presne známy. Človek je však všeobecne málo preskúmaný tvor a je odvážne tvrdiť, že je v jeho konštrukcii niečo nadbytočné.

Stručný prehľad o genóme udáva parameter nazývaný obsah GC (angl. *GC-content*), čiže percentuálny podiel báz G a C k celkovej dĺžke genómu. Vypočíta sa podľa vzorca:

$$\frac{G+C}{A+G+C+T} \times 100 \ [\%] \tag{2.1}$$

kde A, C, G a T sú počty príslušných nukleotidov.

Obsah GC je veľmi variabilný a odvíja sa od taxonomickej[4] skupiny (pozri str. 22). Sám o sebe tento parameter neudáva nič zaujímavé, no význam nadobúda napr. pri porovnávaní dvoch rôznych genómov.

2.3 Baktérie a ich štruktúra

Baktérie sú jednoduché jednobunkové prokaryotické organizmy. Ich rozmery sú rozmanité a pohybujú sa okolo rádu niekoľkých mikrometrov. Sú najrozšírenejším druhom organizmov na svete [23]. Nachádzajú sa vo vode, v pôde, v symbióze s inými živými organizmami, no niektoré sú dokonca schopné existovať aj vo vesmíre, čiže v relatívnom vákuu [7]. V jednom grame pôdy žije asi 40 miliónov baktérií a v jednom mililitri sladkej vody je ich približne milión [22]. Baktérie teda vyžadujú rozmanité podmienky pre život. Považujú sa za najstaršie živé organizmy (až 3 miliardy rokov staré), ktoré sa vyskytujú vo forme fosílií [20].

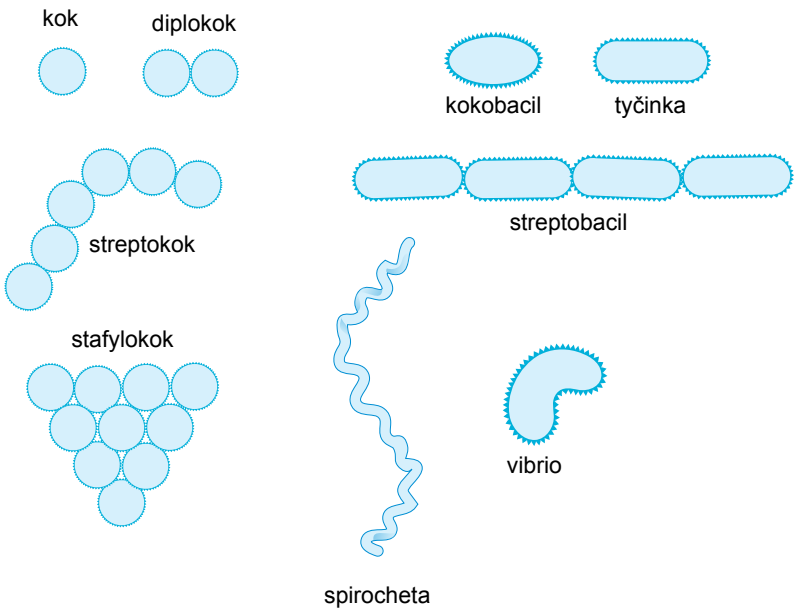
Základné delenie baktérií je podľa tvaru, konkrétne na hlavné skupiny:

- **koky** – majú guľovitý tvar, väčšie druhy tvoria zhluky guľovitých tvarov. Patria sem napr. streptokok, stafylokok.
- **tyčinky (bacily)** – majú predĺžený tvar podobný tyčinkám, väčšie druhy sú takisto tvorené zhlukmi tyčiniek. Príklady: *Escherichia coli*, *Nocardia*.
- **ostatné** – majú rôzne tvary (pretiahnutý, kyjakovitý, špirálovitý atď.), nevytvárajú zhluky. Majú zložitejšiu štruktúru. Patria sem napr. spirochéty, ktoré môžu dosahovať dĺžku až 0,5 mm [18].

Tvary rôznych druhov baktérií sú znázornené na obr. 2.3.

Obdoba jadra v bakteriálnych bunkách sa nazýva nukleoid (viď str. 5). Ich DNA je uložená priamo v cytoplazme väčšinou ako jeden chromozóm s kruhovým tvarom.

[4] taxón – pozri str. 7



Obr. 2.3: Rôzne tvary baktérií [1]

2.3.1 Rozdelenie baktérií

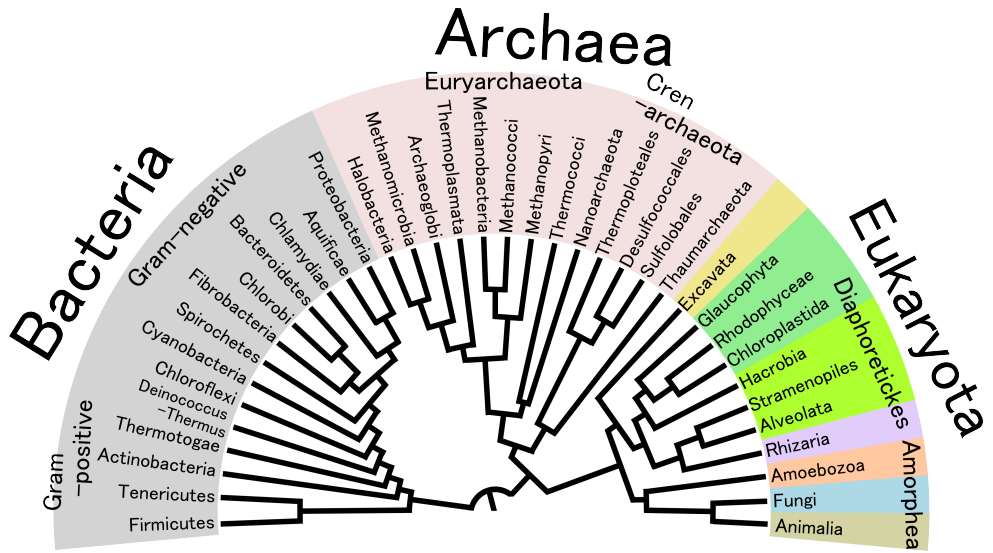
Organizmy na svete sú veľmi rozmanité, napr. baktéria a človek sa nielen geneticky extrémne líšia. Preto bol zavedený systém klasifikácie organizmov do skupín – taxónov. Skúmaním taxónov sa zaoberá vedný obor – taxonómia.

Zmysel taxonómie spočíva v stromovej štruktúre znázorňujúcej hierarchiu živých organizmov, ktorá sa nazýva fylogenetický strom. Jedná sa o n-árnu stromovú štruktúru, ktorá združuje organizmy podľa vývojovej línie.

Obecne sa za život na Zemi považujú organizmy, ktorých stavebnými jednotkami sú aminokyseliny. Tie sa rozdeľujú do 3 hlavných kategórií: baktérie, archaea a eukaryoty. Archaea spolu s baktériami pôvodne tvorili skupinu prokaryotických organizmov, oddelili sa až v roku 1977 [27]. Vtedy boli tieto skupiny pomenované ako Archebacteria a Eubacteria, keďže pôvodne existovala len skupina – prokaryoty. Archaea sa svojimi vlastnosťami nachádza medzi baktériami a eukaryotmi. Vyznačujú sa extrémnymi podmienkami existencie ako veľké koncentrácie NaCl, síry či nízke pH [16].

Baktérie je možné rozdeľovať aj empirickou metódu farbenia podľa Grama. Táto metóda rozdeľuje druhy baktérií na gram-pozitívne (ďalej G+) a gram-negatívne (ďalej G-). Vo výsledku sa pri experimentoch G+ druhy sfarbia do modro-fialova a G- do ružova. Závisí to od vlastností bunkovej steny, ktorú majú G- baktérie zloženú z dvoch vrstiev a celkovo má inú štruktúru. Preto sú z veľkej časti patogény (spôsobujú infekčné choroby). Príslušné kmene oboch skupín sú vyznačené na obr. 2.4.

Štruktúra vývoja alebo aj záznam vo fylogenetickom strome je znázornený v tabuľke 2.1. Bunka „Život“ je koreňom fylogenetického stromu. Ľavý stĺpec smerom zhora nadol konkretizuje skupiny organizmov. Každá nižšia podskupina združuje organizmy s niektorými spoločnými vlastnosťami. Toto delenie nie je však 100% konečné, pretože kvôli skúmaniu mnohých organizmov jedného druhu sa môžu znovu vyčleňovať ďalšie rozdiely. Nižším (špecifickejším) taxónom než druh je tzv. poddruh, ktorý sa označuje určitým identifikátorom



Obr. 2.4: Fylogenetický strom – „strom života“ zobrazujúci 3 hlavné vývojové línie: baktérie, archaea a eukaryoty [2].

alebo opakovaním mena druhu (tak ako aj dobre známy *Homo sapiens sapiens* – tzv. „človek dnešného typu“). Poddruhov však môže byť veľké množstvo.

Život		superkategória, všetky živé organizmy
Doména	Bacteria	ostatné: Archaea, Eukaryoty
Ríša	Eubacteria	z histórie, pred vyčlenením Archaea ako domény
Kmeň	Proteobacteria	zvyčajne podľa životných podmienok
Trieda	Gammaproteobacteria	združuje príbuzné rady
Rad	Enterobacteriales	podrobnejšie rozdelenie
Čeľaď	Enterobacteriaceae	združuje príbuzné rody
Rod	Escherichia	geneticky blízké druhy
Druh	Escherichia coli	konkrétny druh s určitým genómom

Tabuľka 2.1: Popis taxonomických kategórií s príkladom baktérie *Escherichia coli*.

2.3.2 Bakteriálny genóm

Vzhľadom na jednoduchú štruktúru baktérií sú bakteriálne genómy vhodné na analýzu. Preto bol aj prvý prečítaný genóm práve bakteriálny, konkrétne genóm baktérie *Haemophilus influenzae* sekvenovaný v polovici 90. rokov dvadsiateho storočia. Ten je relatívne krátky – má dĺžku asi 1,8 Mbp.

Jednou z najlepšie preskúmaných baktérií je *Escherichia coli*, ktorá sa využíva ako vzorová baktéria v biotechnológiach a genetickom inžinierstve [16]. Je to gram-negatívna tyčinkovitá baktéria, ktorá je súčasťou črevnej mikroflóry teplokrvných živočíchov vrátane človeka. Tam napomáha tvorbe niektorých vitamínov. Pohybuje sa pomocou bičíkov. Jej genóm má veľkosť asi 4,6 Mbp.

Gény u baktérií tvoria viac než 80% dĺžky genómu[5]. Taktiež sa u baktérií nenachádzajú intróny (typické pre eukaryotické organizmy). Štúdie ukazujú, že niektoré baktérie majú kratšie genómy ako mali ich predchodcovia [15].

Do roku 2012 bolo kompletne sekvenovaných 1017 bakteriálnych, 110 archaea a 36 eukaryotických genómov [13]. Eukaryotické genómy sú väčšinou pre svoju zložitosť nedokončené.

[5]U človeka sú to len asi 2%.

Kapitola 3

Úvod do bioinformatiky

S prudkým technologickým vývojom sa do popredia dostáva tiež vedná disciplína s názvom bioinformatika. Bioinformatika v skratke je fúziou molekulárnej biológie a informatiky. Rozvoj informatiky mnohonásobne rozširuje možnosti práce s rozsiahlymi databázami údajov. Pôvodne sa tieto databázy udržiavali v podobe papierových kartoték, no tento spôsob veľmi obmedzoval prístup k údajom a taktiež ich modifikácie. Prínos informatiky je revolučný aj vďaka internetu, ktorý tieto dáta zbavuje pripútanosti ku konkrétnej miestnosti a umožňuje k nim prístup skadekoľvek na svete.

Podľa definície je teda bioinformatika obor, ktorý sa zaoberá najmä spracovaním, prehľadávaním a analýzou dát o sekvencii a o štruktúre biologických makromolekúl [4]. Jednoduchšie povedané je to „využitie počítačov k hľadaniu odpovedí na biologické otázky“ vrátane štatistického spracovania získaných výsledkov [4].

3.1 Biologické databázy

Biologické dáta tvoria masívnu odbornú dátovú množinu. Je, takpovediac, nekonečná, pretože neustále sa vďaka vylepšeným technológiám získavajú nové dáta a takisto sa objavujú nové (presnejšie doteraz neobjavené) organizmy a tento „strom života“ sa neustále rozvetvuje.

Typmi biologických dát môžu byť sekvencie DNA molekúl či proteínov alebo štruktúry molekúl. Taktiež aj informácie o biochemických dráhach – rôznych metabolických procesoch alebo fotosyntéze. Tieto dáta pomáhajú napr. pri vývoji liekov alebo pri skúmaní vzťahov medzi organizmami [24].

Najväčšie sú databázy sekvencií DNA, RNA (ribonukleová kyselina) a proteínov. Keďže je týchto údajov veľké množstvo, bolo v minulosti technicky i časovo náročné ich získavať. Metóda získavania (čítania) sekvencie sa nazýva sekvenovanie DNA. Metód sekvenovania DNA je dodnes známych už mnoho.

Pre správu databáz sekvencií DNA bola vytvorená medzinárodná spolupráca INSDC[1]. Tá spája najväčšie strediská DNA dát v rôznych častiach sveta. Zastrešuje 3 najväčšie databázy GenBank, ENA a DDBJ (tab. 3.1).

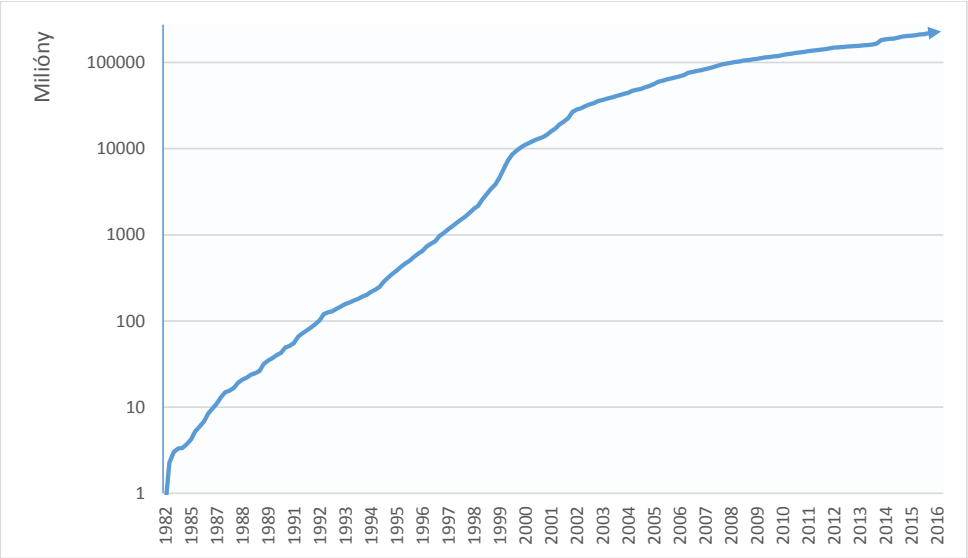
Tieto databázy sú synchronizované – obsahujú rovnaké dáta a zmeny si v pravidelných intervaloch vymieňajú, čím udržujú konzistenciu a aktuálnosť. Takáto paralelizácia zabezpečuje zálohovanie a rozkladá prístupovú záťaž.

[1] International Nucleotide Sequence Database Collaboration

	GenBank	EMBL (ENA)	DDBJ
Sídlo	USA	Európa (Spojené kráľovstvo)	Japonsko
Organizácia	National Center for Biotechnology Information	European Bioinformatics Institute	National Institute of Genetics
Vznik	1982	1974	1986

Tabuľka 3.1: Členské databázy INSDC.

Záznamy v databázach neustále exponenciálne narastajú. NCBI uvádza, že od roku 1982 dodnes sa počet báz zdvojnásobí približne každých 18 mesiacov [12]. Do decembra roku 2016 GenBank zaznamenal vyše 200 miliárd báz tvoriacich asi 200 miliónov sekvencií [11].



Obr. 3.1: Vývoj dát databázy GenBank za posledných 35 rokov.

Databázy sú verejne prístupné. Prístup umožňujú internetové stránky, ktoré sú usporiadané tak, aby umožňovali čo najjednoduchšie prehliadanie alebo sťahovanie.

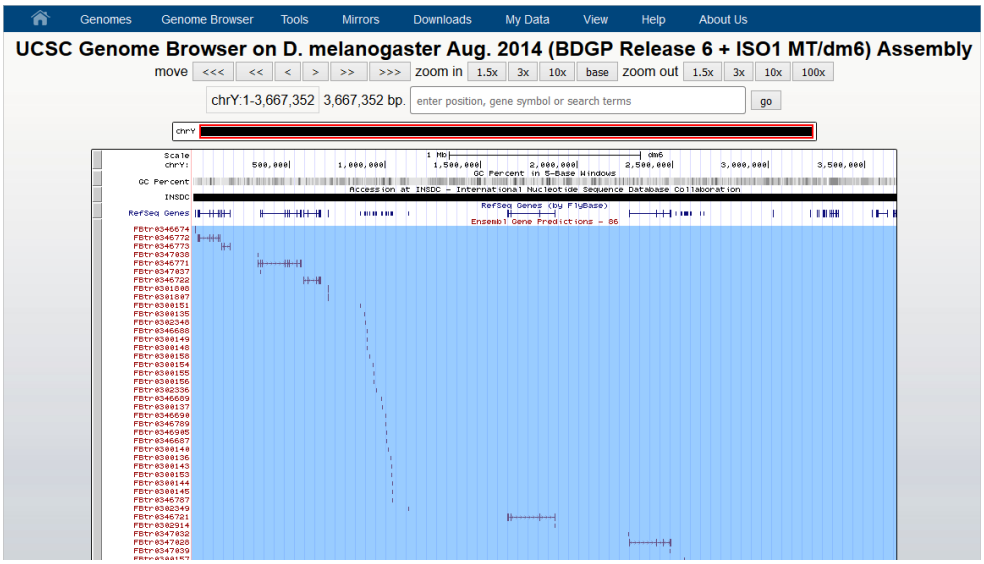
Existujú nástroje, ktoré umožňujú špecifickejšie vyhľadávanie. K hľadaniu genómových dát slúžia tzv. genómové prehliadače (angl. *genome browser*). Štruktúrou sú to tiež databázy, no dáta čerpajú väčšinou z hlavných databáz. Umožňujú podrobnú vizualizáciu na rozdiel od holých sekvencií. Genome browsery sú dobre optimalizované a dokážu zobraziť sekvenciu vrátane zvýraznených génov či iných detailov.

Jedným z takýchto prehliadačov je **UCSC Genome Browser**[2], ktorý vyvíja University of California, Santa Cruz (UCSC). Po vyhľadaní organizmu poskytuje prehľad všetkých, k nemu prislúchajúcich, záznamov, najmä chromozómy a plazmidy. Tieto záznamy sú verejné a odkazujú aj do iných databáz. Anotácie sekvencií sú odvodené zo záznamov z GenBank.

[2] http://genome.ucsc.edu/

Vizualizácia sekvencie je na zobrazená na obr. 3.2. Zobrazený je chromozóm Y vínnej mušky. Panel s navigáciou v hornej časti umožňuje posúvanie alebo zväčšovanie sekvencie. Sekvencia je zobrazená horizontálne. Údaje zobrazované v grafe sú voliteľné, čiže je možné zobraziť len to, čo je potrebné. Roztrieštené úseky v dolnej časti označujú jednotlivé gény a ich pozíciu v rámci sekvencie.

Prehliadač taktiež poskytuje aj zobrazenie 3D štruktúr proteínov, na ktoré využíva údaje z databáz proteínov, napr. PDB (Protein Data Bank).



Obr. 3.2: Užívateľské rozhranie UCSC Genome Browseru.

Ďalším webovým nástrojom k prehliadaniu sekvencií je **Ensembl genome database project**. Založený bol výskumným centrom European Bioinformatics Institute (EBI), ktoré je súčasťou európskeho inštitútu EMBL. O 10 rokov od svojho vzniku, roku 2009, vznikol sesterský projekt Ensembl Genomes špecializovaný na bezstavovce, tj. baktérie, protista[3], huby a rastliny [5].

Vizualizácia ponúka zväčša rovnaké možnosti ako UCSC Genome Browser, no javí sa prehľadnejšie a vyžaduje menšiu hardvérovú náročnosť. Umožňuje aj vyhľadávanie na základe génu, ktoré vyhľadá organizmy, ktoré daný gén obsahujú.

Svoj prehliadač má aj GenBank. Ten obsahuje veľké množstvo dát, ale ponúka podobné funkcie ako predchádzajúce alternatívy.

Za zmienku stojí aj databáza **RefSeq** (Reference Sequence), ktorú (takisto ako GenBank) spravuje NCBI. Táto databáza obsahuje vždy len jeden záznam z určitej molekuly (DNA, RNA, proteínu). Ostatné databázy väčšinou obsahujú viacero verzií sekvencie [17].

Najzaujímavejšou funkciou väčšiny prehliadačov je možnosť sťahovať dáta. U DNA sekvencií sa však jedná o objemné balíky dát.

3.2 Formáty biologických dát

Mnoho druhov biologických dát je pre počítače nutné ukladať serializovane – za sebou do súborov. Pri sekvenciách DNA nukleotidy nasledujú bezprostredne za sebou, no uchovávať

[3] súhrnné označenie pre jednobunkové eukaryotické organizmy

napr. 3D štruktúry proteínov nie je jednoduché. Databázy pre tento účel musia voliť vhodný spôsob ukladania.

Sekvencie DNA či RNA sú v databázach uložené lineárne. Rozdeľujú sa len na miestach, kde sú hranice chromozómov. Pri prokaryotických organizmoch s často jedným chromozómom sú sekvencie v celku a jednoducho sa s nimi pracuje.

Dodatočné informácie o sekvenciách (metadáta) bývajú dostupné v iných – štrúkturovaných formátoch. Tie obsahujú pozície génov a ich hierarchiu. Najmä pri eukaryotických organizmoch môžu jednotlivé úseky tvoriť viacero kombinácií – potenciálnych proteínov. Takýto proces sa nazýva alternatívny zostrih či z angl. splicing.

Jedným z najjednoduchších formátov pre uloženie sekvencií nukleotidov či proteínov je textový formát **FASTA** (obr. 3.3).

```
>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
...
```

Obr. 3.3: Ukážka formátu FASTA.

Na prvom riadku formátu FASTA, ktorý začína symbolom `>`, sú základné informácie. V tomto prípade je zvolený predpis podľa databázy RefSeq – `NC` značí kompletný genóm, nasleduje identifikátor organizmu a prípadné doplnkové informácie (či sa jedná o kompletný genóm alebo len jeden chromozóm apod.). Od druhého riadka do konca súboru nasleduje samotná sekvencia nukleotidov. Databázy rozdeľujú sekvenciu na konštantnú dĺžku riadka, v prípade vzorky na obr. 3.3, ktorá je získaná prostredníctvom NCBI je to 70 znakov na riadok. Ensembl sekvencie rozdeľuje na 60 znakov na riadok.

Tento formát je v podstate najintuitívnejší – ide o „sekvenciu písmen“. Keďže nukleotidy sú 4, pri ukladaní do počítača v binárnej sústave je možné pre zmenšenie objemu dát využiť kódovanie nukleotidov na 2 bity. U proteínov je to zložitejšie, keďže aminokyselín je viac než 20, ale aj tak je to úspornejšie, pretože sa využije asi len 5 bitov na rozdiel od písmen, ktoré sú kódované na 7 a viac bitov.

K anotácii sekvencií sa využíva viacero štruktúrovaných formátov, ktoré sa môžu vyskytovať v súbore so sekvenciou alebo samostatne. Tieto formáty sa líšia z väčšej časti len syntaxou. Vždy ale obsahujú začiatočné a koncové pozície úsekov a informácie, čo ten– ktorý úsek značí. Bežná štruktúra formátov, kde je sekvencia spolu s popisom, je v tvare: informácie o sekvencii – anotácia úsekov – sekvencia. Medzi známe formáty patria napr.:

GenBank
Formát databázy GenBank, obsahuje podrobné informácie o sekvencii, vrátane autorov, fylogenézy (vývojovej línie) organizmov či dátumov a autorov rôznych úprav. Formát je rozdelený na 2 stĺpce. V ľavom stĺpci je typ údaju a v pravom stĺpci hodnota. Sekvencia sa nachádza za kľúčovým slovom `ORIGIN`.

GFF3 (Generic Feature Format)
Ide o 9-stĺpcový textový formát, kde stĺpce sú oddelené tabulátormi (obr. 3.4). Vďaka

tomu je formát vhodný pre shellové príkazy (napr. `grep`) [19]. Jednotlivé riadky popisujú konkrétne záznamy o častiach sekvencie. Sekvencia znakov `##` značí direktívu, `#` komentár.

Význam stĺpcov je nasledovný:

1. **stĺpec – *seqid*:** Identifikátor sekvencie, ku ktorej jednotlivé záznamy patria.
2. **stĺpec – *source*:** Zdroj sekvencie, väčšinou algoritmus, ktorý sekvenciu vyprodukoval alebo databáza, z ktorej sekvencia pochádza.
3. **stĺpec – *type*:** Typ záznamu (exón, gén, mRNA apod.).
4. **a 5. stĺpec – *start* a *end*:** Celé čísla určujúce začiatok a koniec označovanej časti sekvencie. Platí, že $start \leq end$. Rovnosť sa využíva pri vložení doplňujúcich informácií.
6. **stĺpec – *score*:** Desatinné číslo udávajúce ohodnotenie záznamu. Hodnoty sú voliteľné.
7. **stĺpec – *strand*:** Orientácia vlákna, kladná (+) alebo záporná (−) orientácia. V prípade nevláknitých sekvencií je hodnota . (bodka).
8. **stĺpec – *phase*:** Číslo, ktoré vyjadruje počet báz nutných pre posunutie k najbližšiemu kodónu.
9. **stĺpec – *attributes*:** Zoznam atribútov sekvencie vo formáte `názov=hodnota` oddelené bodkočiarkami. Niektoré atribúty sú:
   - ID – unikátny identifikátor záznamu
   - Name – názov záznamu zobrazovaný užívateľovi, nie je unikátny
   - Parent – rodičovský záznam konkrétneho záznamu, umožňuje radiť záznamy do hierarchických štruktúr. Jeden záznam môže mať viacero rodičov

Na obr. 3.4 je zobrazený popis sekvencie s názvom `abc123`. Prvý riadok označuje verziu GFF, druhý riadok identifikátor a interval dĺžky sekvencie. Tretí riadok určuje, že v intervale 1 000 až 9 000 sa nachádza gén s identifikátorom `gene01`. Riadky 3 a 4 sú úseky mRNA (`mRNA01` a `mRNA02`), ktoré už tvoria stromovú štruktúru – gén `gene01` je ich rodičovský prvok. Piaty riadok značí úsek (exón), ktorý je naviazaný na `mRNA02`. Jeden úsek môže byť zahrnutý do viacerých spojení (splicing). Toto je vylepšenie oproti staršej verzii GFF2.

```
##gff-version 3
##sequence-region abc123 1 150000
abc123 . gene 1000 9000 . + . ID=gene01;Name=G1
abc123 . mRNA 1050 9000 . + . ID=mRNA01;Parent=gene01
abc123 . mRNA 1300 9000 . + . ID=mRNA02;Parent=gene01
abc123 . exon 1300 1500 . + . ID=exon01;Parent=mRNA02
...
```

Obr. 3.4: Ukážka formátu GFF3.

Kompletná špecifikácia formátu GFF3 je na [19].

Ďalšie používané formáty sú tiež: **BED** (Browser Extensible Data), **XML**, **ASN.1** a iné.

3.3 Popis DNA sekvencií

Sekvencie (nielen) DNA slúžia k analýze organizmov. Každý organizmus má špecifickú DNA a tak je možné jednotlivé DNA porovnávať, čím sa zisťujú rôzne vzťahy medzi organizmami.

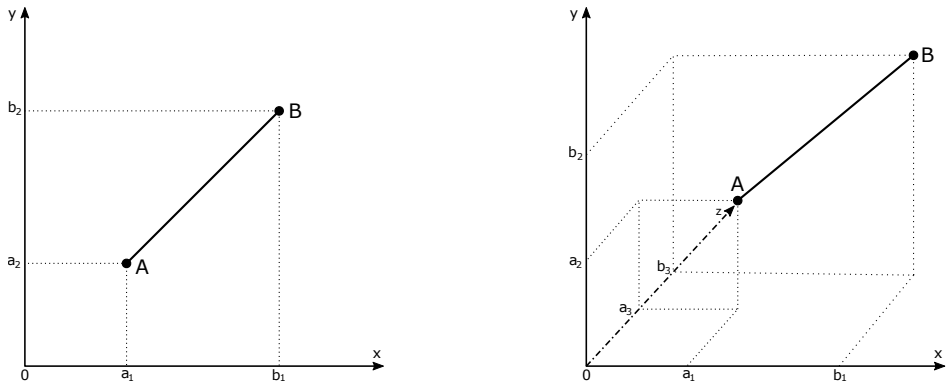
Samotná sekvencia je dlhý reťazec zložený zo štyroch písmen (nukleotidov). Preto nie je veľa možností ich spracovania. Dajú sa u nich sledovať len jednoduché vlastnosti, napr. ako sú nukleotidy za sebou usporiadané alebo ako husto sú na danej dĺžke rozmiestnené. Ďalej je možné skúmať n-tice nukleotidov (dvojice, trojice atď.) a ich počty. Tieto parametre už udávajú presnejšie údaje o sekvencii. Obecne platí, že čím dlhšia je sekvencia, tým presnejší je to ukazovateľ.

Každú sekvenciu je možné rozložiť na jej základné prvky – nukleotidy. Získame tak počty nukleotidov. Takto sa určuje aj základný parameter sekvencie GC-content. Rovnakým spôsobom sa dajú určiť aj počty dvojíc, trojíc či iných n-tíc. Vypočítaním týchto výskytov získame usporiadanú k-ticu, kde každý prvok je pár (n-tica nukleotidov – počet jej výskytov). Ak budeme brať do úvahy dvojice až pätice nukleotidov, získame pre sekvenciu napríklad takúto usporiadanú k-ticu:

$$A = \Big((\mathrm{AA}, 8), (\mathrm{AC}, 10), (\mathrm{AG}, 9), \ldots, (\mathrm{TTTTT}, 2)\Big) \tag{3.1}$$

ktorá združuje atribúty pôvodnej sekvencie. Ak v tomto prípade vezmeme do úvahy všetky dvojice až pätice nukleotidov, počet prvkov k bude $4^2 + 4^3 + 4^4 + 4^5 = 1360$. Prirodzene platí, že čím je n-tica nukleotidov dlhšia, tým menší bude počet jej výskytov.

Takáto usporiadaná k-tica pripomína jednoduchý bod v k-rozmernom priestore. Preto sa porovnávanie sekvencií dá riešiť aj týmto spôsobom. Porovnanie dvoch sekvencií ako bodov tvorených ich atribútmi teda závisí od pozície týchto bodov v k-rozmernom priestore (obr. 3.5).



(a) 2D priestor (rovina)

(b) 3D priestor

Obr. 3.5: Vzdialenosť dvoch bodov.

3.4 Vzdialenosť dvoch DNA sekvencií

Vzdialenosť medzi dvomi bodmi musí byť vždy nezáporná. Čím je vzdialenosť dvoch bodov menšia (bližšia k nule), tým viac sa body blížia k totožnosti a nimi popísané sekvencie sú si podobnejšie.

Ak máme dva body A a B, pre vzdialenosť medzi nimi platia nasledovné pravidlá [6]:

- $\mathrm{d}(A, B) \geq 0$
- $\mathrm{d}(A, A) = \mathrm{d}(B, B) = 0$
- $\mathrm{d}(A, B) = \mathrm{d}(B, A)$

Vzdialenosť dvoch bodov sa dá určiť viacerými spôsobmi. Základný a najpoužívanejší spôsob je **Euklidovská vzdialenosť**.

Euklidovská vzdialenosť dvoch bodov je založená na báze Pytagorovej vety. Jedná sa v podstate o dĺžku prepony v pravouhlom trojuholníku, kde odvesny sú rozdiely jednotlivých súradníc. Pre dvojrozmerný priestor je definovaná vzťahom:

$$\mathrm{d}(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \tag{3.2}$$

kde $A = (a_1, a_2) \in \mathbb{R}^2$ a $B = (b_1, b_2) \in \mathbb{R}^2$ sú body v dvojrozmernom priestore.

Analogicky pravidlo platí aj pre n-rozmerný priestor:

$$\mathrm{d}(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{3.3}$$

kde $A = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ a $B = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ sú body v n-rozmernom priestore.

Ďalšou možnosťou je napr. tzv. **Hammingova/Manhattanská vzdialenosť**. Obidve vzdialenosti majú rovnaký princíp. Hammingova vzdialenosť sa využíva v informatike pri detekcii a oprave chýb na úrovni bitov [10]. Obe určujú počet zmien v rámci dvoch súborov. Ich definícia je nasledovná:

$$\mathrm{d}(A, B) = \sum_{i=1}^{n} |a_i - b_i| \tag{3.4}$$

Tieto metódy vracajú reálne číslo v intervale $\langle 0; \infty \rangle$, pričom 0 znamená zhodu, v opačnom prípade vyššie číslo vyjadruje väčšiu rozdielnosť. Preto metódy môžeme nazvať ako miery nepodobnosti alebo miery odlišnosti. Euklidovská vzdialenosť vyjadruje priamu vzdialenosť („vzdušnú čiaru“) a Manhattanská vzdialenosť akumuluje len postupné zmeny. Preto je euklidovská vzdialenosť väčšinou presnejšia, čo potvrdzuje aj nasledujúci príklad:

$$A = (2,\ 3,\ 2{,}6) \quad B = (2{,}8,\ 2{,}3,\ 2{,}7) \tag{3.5}$$

$$\mathrm{E}(A, B) = 1{,}07 \quad \mathrm{M}(A, B) = 1{,}6 \tag{3.6}$$

kde E je euklidovská vzdialenosť a M je manhattanská vzdialenosť.

Existujú aj iné metódy, ktoré určujú mieru podobnosti – korelácie. Patrí sem napr. Pearsonov korelačný koeficient, ktorý nadobúda hodnoty v intervale $\langle -1; 1 \rangle$. Nulová hodnota značí nepodobnosť, jednotková hodnota znamená úplnú zhodu a hodnota -1 určuje antikoreláciu (opačnú zhodu). Antikorelácia nastáva v prípadoch, ak máme napr. jednu postupnosť čísel $1, 2, 3$ a druhú $3, 2, 1$. Medzi týmito dvomi postupnosťami je maximálna antikorelácia.

Pri viacerých bodoch, kedy nie je vhodné použiť vzdialenosť, sa využívajú rôzne štatistické metódy. K určeniu variability množiny bodov x sa využíva napr. metóda smerodajná odchýlka s_x:

$$s_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N}} \tag{3.7}$$

kde $\overline{x}$ je aritmetický priemer danej množiny. N je počet prvkov množiny. Táto veličina udáva priemernú hodnotu, okolo ktorej sa hodnoty množiny rozprestierajú.

Často sa používa aj veličina variačný koeficient v_x, ktorá normalizuje danú smerodajnú odchýlku k aritmetickému priemeru prvkov množiny hodnôt. Používa sa napr. vtedy, keď chceme porovnať dve množiny hodnôt s rozdielnymi úrovňami hodnôt (napr. desiatky a tisíce). Obvykle sa vyjadruje v percentách:

$$v_x = \frac{s_x}{\overline{x}} \cdot 100 \ [\%] \tag{3.8}$$

3.5 Zhlukovanie DNA sekvencií

Ak máme dve rôzne skupiny atribútov, môžeme z nich odvodiť jednu skupinu, ktorá združuje ich vlastnosti. Tento proces sa nazýva zhlukovanie. Takto je možné zhlukovať napr. farby vo farebnom modeli RGB. Z červenej a zelenej sa vytvorí žltá farba. Žltá farba je jednotná, avšak vo svojej podstate je tvorená pôvodnými dvomi farbami.

Zhlukovanie je možné tiež využiť pri tvorbe taxonomických jednotiek a všeobecne pri tvorbe fylogenetického stromu, keďže ten združuje podobné organizmy. Typ zhlukovania na báze spájania prvkov do hierarchických štruktúr sa nazýva hierarchické zhlukovanie. Na rozdiel od nehierarchického zhlukovania, v tomto prípade nie sú vopred známe skupiny vytvorených zhlukov [9].

Mnohé metódy založené na posúvaní a následnom porovnávaní DNA sekvencií sú však výpočtovo náročné, preto sa využívajú metódy, ktoré umožňujú nájsť suboptimálne riešenie v prijateľnom čase [3].

Hierarchické metódy sú vhodné pre zhlukovanie DNA sekvencií. Cieľom je vytvoriť skupiny – zhluky, ktoré reprezentujú taxonomické kategórie. Tieto zhluky je nutné ďalej spájať do finálnej stromovej štruktúry. Problémom je však to, ako určiť podobnosť dvoch zhlukov.

Ak sú sekvencie popísané pomocou počtov n-tíc nukleotidov, atribúty sú číselné hodnoty a teda je pre ich zhlukovanie možné využiť rôzne metódy priemerovania ako aritmetický, geometrický alebo kvadratický priemer.

Ak máme dve usporiadané n-tice frekvencií nukleotidov A a B:

$$\begin{aligned} A &= \big((\mathrm{AA}, 8), (\mathrm{AC}, 10)\big) \\ B &= \big((\mathrm{AA}, 4), (\mathrm{AC}, 12)\big) \end{aligned} \tag{3.9}$$

môžeme z nich vytvoriť n-ticu C, ktorá tieto dve zhlukuje. Ak použijeme aritmetický priemer, výsledná n-tica C bude nasledovná:

$$C = \big((\mathrm{AA}, \frac{8+4}{2}), (\mathrm{AC}, \frac{10+12}{2})\big) = \big((\mathrm{AA}, 6), (\mathrm{AC}, 11)\big) \tag{3.10}$$

Výsledná usporiadaná n-tica C je teda tvorená rovnakými atribútmi ako A a B. Je teda možné ju znovu zhlukovať.

Kapitola 4

Návrh a implementácia nástrojov

Podstata práce spočíva v skúmaní DNA sekvencií bakteriálnych genómov. Úlohou je teda vytvoriť nástroj, ktorý umožní analýzu získaných DNA sekvencií na základe počtov rôznych n-tíc nukleotidov.

4.1 Implementačné detaily

Vzhľadom na neustály vývoj v oblasti informačných technológií existuje v dnešnej dobe mnoho komponentov (knižnice programovacích jazykov, frameworky) pre prácu s biologickými dátami. Tieto nástroje sú vyvíjané dôkladne, no stále nútia človeka osvojovať si nové prostredia, ktoré sú síce kvalitné, no poskytujú množstvo, pre konkrétnu situáciu, nepotrebných funkcií.

V rámci experimentovania pri tejto práci budú stačiť základné matematické a štatistické funkcie, ktoré sú jednoducho implementovateľné. Bodom zadania je vytvoriť nástroj, ktorý umožní analýzu počtov rôznych n-tíc nukleotidov. Keďže táto funkcia je dobre dekomponovateľná, je vhodné vyčleniť operácie s dátami do samostatných modulov.

Zvolený bol spôsob, kedy sa jednotlivé skripty spúšťajú samostatne, tj. neexistuje globálny bod prístupu, z ktorého sú jednotlivé operácie volené. Tento postup bol zvolený najmä kvôli prehľadnosti, aj keď pre nezainteresovaného človeka to môže spočiatku pôsobiť chaoticky.

Za výkonné nástroje boli zvolené programové skripty v jazykoch `bash` a `python` verzie 2.7 kvôli svojej rýchlosti. Ako operačný systém bol zvolený Linux. Jazyk `python` bol zvolený najmä kvôli svojej jednoduchosti práce so štruktúrovanými dátami. Umožňuje efektívnu prácu so zoznamami, čo bolo využité pri práci s načítaním sekvencií.

Ovládanie skriptov je riešené možno najprimitívnejším, no niekedy najefektívnejším spôsobom, cez príkazový riadok (na operačných systémoch unixového typu – `shell`). Dôvodom tohoto prístupu je, že nie je potrebné počas behu programu doň explicitne zasahovať. Vzhľadom na objem spracovávaných dát sú výpočtové operácie časovo relatívne náročné a údaje sú generované v širších časových intervaloch. Tým pádom sa toto prostredie s priebežným výpisom stavov do súborov javí ako vhodné.

Nástroj teda musí pracovať s textovými formátmi dát FASTA a GFF3 pre metadáta sekvencií. Musí byť schopný tieto dáta „rozparsovať“ na jednotlivé frekvencie nukleotidov. S frekvenciami je možné vykonávať operácie ako porovnávanie dvoch zoznamov alebo zhlukovanie dvoch zoznamov do jedného.

Prevažná časť skriptov tvoriacich analyzačný nástroj je nasledovná:

comp.py – porovná dva vstupné súbory obsahujúce frekvencie n-tíc a vypočíta koeficient ich podobnosti. Za metódu rozpoznania bola zvolená **Euklidovská vzdialenosť**. Dĺžka n-tíc braných v úvahu sa nastavuje ako parameter.

dismantle.py – rozloží vstupnú sekvenciu na jednotlivé frekvencie nukleotidov. Prijíma formáty FASTA a prípadne anotáciu vo formáte GFF3. Tiež umožňuje pre rozklad zvoliť konkrétny podinterval sekvencie alebo normalizovať počty výskytov n-tíc k celkovej dĺžke sekvencie.

comutate.py – na vstup prijíma dva súbory s počtami n-tíc a skript z nich na základe aritmetického priemeru metódou zhlukovania vytvorí jeden zoznam.

reduce.py – zo súhrnnej tabuľky frekvencií vypočíta variabilitu n-tíc a výsledky zoradí (nájde „najkolísavejšie“ n-tice). Umožňuje zvoliť percentuálny počet najvariabilnejších n-tíc.

tools.py – obsahuje pomocné funkcie (ukončenie s chybou apod.).

analyze.sh – rozloží všetky sekvencie z daného adresára na frekvencie n-tíc a vyprodukuje súhrnnú tabuľku frekvencií, ktorú následne využíva skript `reduce.py`.

classifier.sh – porovná podsekvencie nastavených dĺžok so vzorovými dátami cez zvolený rozsah dĺžok n-tíc. Produkuje percentuálne úspešnosti rozpoznania k jednotlivým vzorkám.

heat_map.sh – spracuje údaje zo skriptu `seq.sh` a zostaví tabuľku podobností jednotlivých fragmentov sekvencií.

maketree.sh – z adresára obsahujúceho súbory s frekvenciami n-tíc vytvorí na báze zhlukovania (skript `comutate.py`) stromovú štruktúru podobností (fylogenetický strom).

multicomp.sh – porovná jeden zoznam frekvencií s N ďalšími a zoradí výsledky podľa podobnosti. Zastrešuje skript `comp.py` a je využívaný vo všetkých porovnávacích operáciách.

multijoin.sh – rozširuje funkciu unixového príkazu `join` na základe zhody prvého stĺpca s oddeľovačom „tabulátor“.

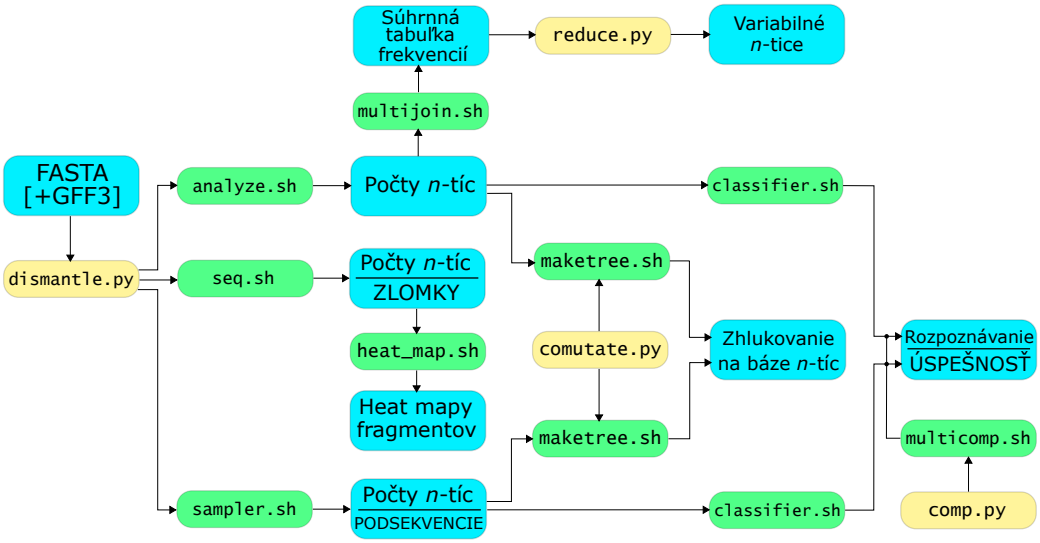
sampler.sh – generuje pseudonáhodné podsekvencie zo vzorových sekvencií podľa zvolených dĺžok.

seq.sh – rozloží všetky sekvencie na zvolený počet fragmentov – zlomkov a tie rozloží na určité n-tice.

tableize.sh – po dokončení rozpoznania vytvorí výsledkovú tabuľku ako je napr. 5.2.

Pythonové skripty (`*.py`) tvoria väčšinou jadro shellových skriptov (`*.sh`), pričom užívateľ môže nastavovať rôzne parametre spúšťania. Je dôležité poradie spúšťaní skriptov – skript `dismantle.py` najprv rozloží vzorky na frekvencie n-tíc a až potom s nimi môžu iné skripty pracovať. Skripty samotné obsahujú obecnú nápovedu pre dôkladnejšie ovládanie. Výstup skriptov je väčšinou jednoduchý – v textovej podobe, takže pre príjemnejšiu vizualizáciu je nutné použiť iný program.

Na obr. 4.1 je znázornený diagram dátových tokov jednotlivých modulov. Modrá farba označuje stavy, resp. výsledky predchádzajúcich operácií. Zelenou farbou sú označené shellové skripty, ktoré vo väčšine vykonávajú hlavnú funkciu. Do shellových skriptov sú navedené oranžové pythonové skripty, ktoré plnia konkrétnu funkciu ako rozloženie sekvencií, porovnávanie či zhlukovanie. Počiatočným stavom je stav s označením FASTA [+GFF3], ktorý vyjadruje podmienku vzorových dát vo formátoch FASTA a voliteľne aj GFF3. V diagrame nie sú znázornené skripty `tools.py` a `tableize.sh`. Modul `tools.py` obsahuje len súhrn funkcií, ktoré sú importované ostatnými pythonovými skriptami. Skript `tableize.sh` je využiteľný po dokončení rozpoznávania skriptom `classifier.sh`.



Obr. 4.1: Diagram dátových tokov.

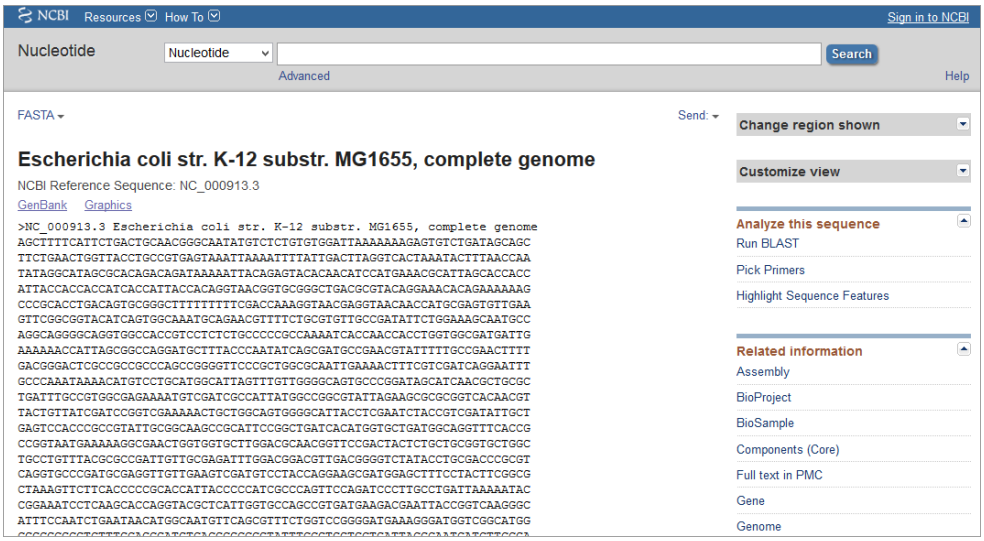
Kapitola 5

Experimentálna časť

Táto kapitola prezentuje hlavnú časť práce. Cieľom experimentov je zistiť, či na základe frekvencií jednotlivých nukleotidov je možné určiť, z ktorej baktérie skúmaná sekvencia pochádza. Čiže hľadáme výraznejšie odlišnosti u genómov rôznych druhov baktérií na základe n-tíc nukleotidov.

5.1 Zvolené vzorky baktérií

K účelu práce bolo nutné získať potrebné dáta z voľne dostupných zdrojov. Základné sekvencie vo formáte FASTA boli získané prostredníctvom genómového browsera UCSC Microbial Genome Browser, ktorý odkazuje na webstránky NCBI, konkrétne do databázy RefSeq (obr. 5.1). Identické dáta obsahuje aj databáza GenBank. K týmto sekvenciám boli tiež stiahnuté anotácie vo formáte GFF3, ktoré budú takisto využité pre skúmanie rozdielnych vlastností voči kompletným genómom.



Obr. 5.1: Webstránka prehliadača NCBI.

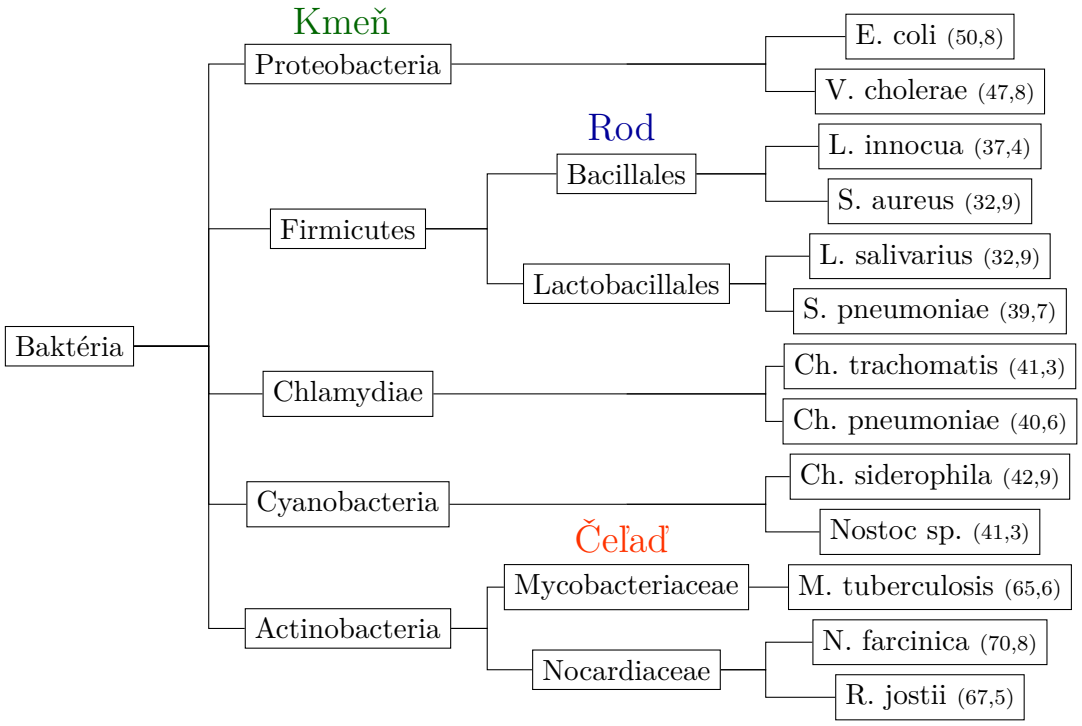
V sekcii 2.3.1 bolo predstavené základné delenie baktérií podľa taxomických kategórií. Výber vzoriek bol koncipovaný tak, aby boli obsiahnuté baktérie z rôznych kmeňov. Dostupných kompletných genómov však nie je mnoho. Preto s ohľadom na dostupnosť bolo za vzorové baktérie vybraných nasledovných 13 baktérií z celkovo piatich kmeňov:

Druh	Kmeň	Rad	G
Escherichia coli	Proteobacteria	Enterobacteriales	−
Chlamydia trachomatis	Chlamydiae	Chlamydiales	−
Chlamydophila pneumoniae	Chlamydiae	Chlamydiales	−
Chroogloeocystis siderophila	Cyanobacteria	Chroococcales	−
Lactobacillus salivarius	Firmicutes	Lactobacillales	+
Listeria innocua	Firmicutes	Bacillales	+
Mycobacterium tuberculosis	Actinobacteria	Actinomycetales	±
Nocardia farcinica	Actinobacteria	Actinomycetales	+
Nostoc sp.	Cyanobacteria	Nostocales	−
Rhodococcus jostii	Actinobacteria	Actinomycetales	+
Staphylococcus aureus	Firmicutes	Bacillales	+
Streptococcus pneumoniae	Firmicutes	Lactobacillales	+
Vibrio cholerae (chromozóm 2)	Proteobacteria	Vibrionales	−

Tabuľka 5.1: Zoznam vzorových baktérií.

K zvoleným baktériám je na internete dostupných viacero genómových sekvencií, ktoré patria rôznym poddruhom. Tie sa navzájom líšia len nepatrne. V tomto prípade nie sú tieto detaily podstatné a nebudú ďalej brané v úvahu. Cieľom bolo vybrať zástupcov rôznych kmeňov.

Na obr. 5.2 je zjednodušene znázornený fylogenetický strom vzorových baktérií. V zátvorkách sú príslušné obsahy GC. Vypočíta ich skript `dismantle.py`.

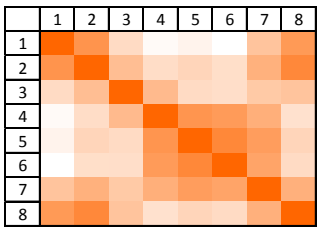


Obr. 5.2: Fylogenetický strom vzorových baktérií.

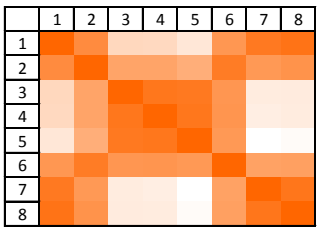
Z obrázku je zrejmé, že taxonomicky bližšie druhy majú podobný obsah GC, ktorý je dosť variabilný. U kmeňa Actinobacteria dosahuje až okolo 70 %.

Aby bolo možné objektívne pracovať so sekvenciami, je treba preskúmať ich zloženie. Ak vystrihneme zo sekvencie určitý úsek, môže sa výrazne líšiť od úseku vystrihnutého z inej časti tej istej sekvencie. Preto môže byť napr. obsah GC na začiatku sekvencie výrazne vyšší ako na jej konci. To je spôsobené najmä kódujúcimi oblasťami, ktoré v rôznych častiach kódujú rôzne proteíny.

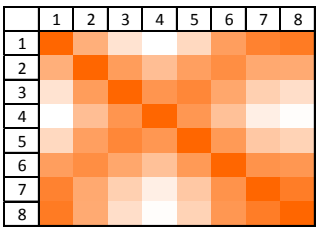
Rozdelenie každej zo vzorových sekvencií na osem fragmentov (osmín) je znázornené na obr. 5.3. Každý obrázok vyjadruje jednu vzorku. Na oboch osách sú nanesené rovnaké údaje a farebné hodnoty v bunkách značia mieru podobnosti dvoch fragmentov. Najtmavšie hodnoty sú na diagonále, kde je fragment porovnávaný so sebou samým a tým pádom má 100% zhodu podobnosti. Bledé miesta vyjadrujú malú podobnosť dvoch príslušných fragmentov. Na porovnanie boli využité trojice nukleotidov a skript `heat_map.sh`.



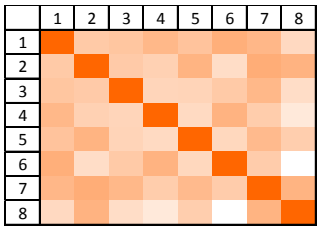
(a) E. coli



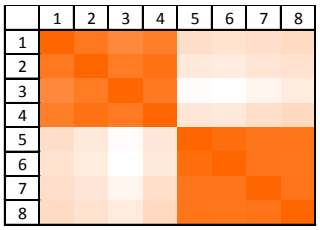
(b) Ch. trachomatis



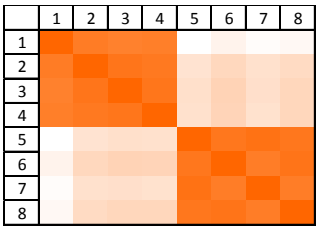
(c) Ch. pneumoniae



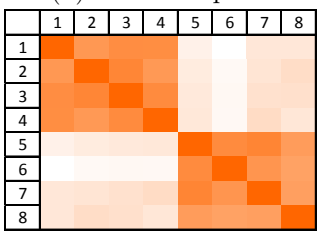
(d) Ch. siderophila



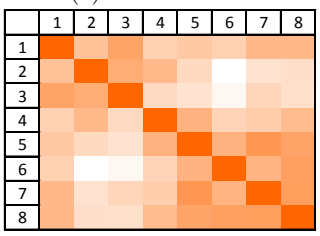
(e) L. salivarius



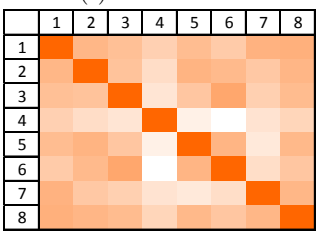
(f) L. innocua



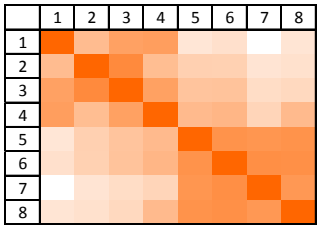
(g) M. tuberculosis



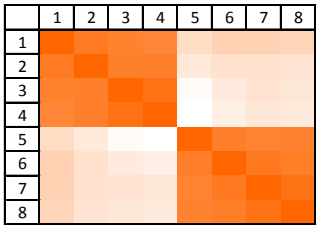
(h) N. farcinica



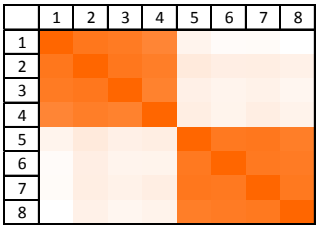
(i) Nostoc sp.



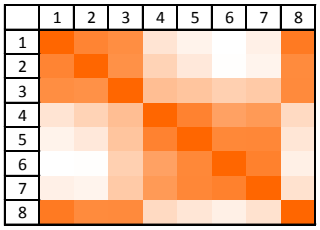
(j) R. jostii



(k) S. aureus



(l) S. pneumoniae



(m) V. cholerae

Obr. 5.3: Heat mapy značiace podobnosť fragmentov vzorových baktérií.

Z týchto zobrazení je možné „od oka“ porovnávať sekvencie, avšak také porovnanie je veľmi nepresné. No obr. (e), (f), (k) a (l) sú vizuálne dosť podobné a podľa stromu 5.2 je to celkom presné určenie, pretože všetky 4 baktérie patria do jedného kmeňa, Firmicutes. Takýto graf značí to, že prvá polovica sekvencie je výrazne odlišná od druhej polovice. Veľkú podobnosť vykazujú aj (b) a (c) alebo (a) a (m).

Presné rozpoznanie určitej sekvencie je teda náročné vzhľadom na veľké množstvo faktorov, ktorých je nutné zohľadniť čo najviac.

5.2 Rozpoznávanie sekvencií

Dĺžky genómov vzorových baktérií sú rôzne. Pohybujú sa približne od 0,38 Mbp do 7,8 Mbp. Preto je zrejmé, že ak vypočítame frekvencie nukleotidov, tak celkové počty budú kolísať vzhľadom k dĺžke sekvencie. Kvôli tomu je vo všetkých operáciách so sekvenciami používaná normalizácia počtu výskytov k celkovej dĺžke sekvencie. Tá je realizovaná vydelením získaného počtu dĺžkou sekvencie a vynásobením 100. Túto funkciu má skript `dismantle.py` s parametrom `-n`.

Príkaz na obr. 5.4 vykoná rozloženie sekvencie zo súboru ecoli.fasta na n-tice nukleotidov, kde $n = 1$. Číslo n udáva hodnota parametra `-l`.

```
$ ./dismantle.py -f bacteria/ecoli.fasta -l 1
A       1142742
C       1180091
G       1177437
T       1141382
```

```
$ ./dismantle.py -f bacteria/ecoli.fasta -l 1 -n
A       24.619295
C       25.423944
G       25.366766
T       24.589995
```

Obr. 5.4: Ukážka normalizácie podľa dĺžky.

Súčet výskytov z hornej časti v tomto prípade prirodzene udáva celkovú dĺžku sekvencie (~4,6 Mbp) a výskyty z dolnej časti zase percentuálne zastúpenie nukleotidov. Všetky počty v dolnej časti sú teda normalizované podľa dĺžky danej sekvencie.

U vyšších n-tíc (dvojice, trojice apod.) sa celkový počet výskytov znižuje. Preto aj normalizované hodnoty budú nízke. To však nie je problém, keďže konkrétna n-tica sa porovnáva s rovnakou n-ticou inej sekvencie.

5.2.1 Rozpoznávanie celých sekvencií

Nasleduje experiment, ktorý spočíva v tom, že zo vzorových sekvencií vystrihneme menšie sekvencie o pevne zvolených dĺžkach. Tieto kratšie sekvencie sa budeme snažiť rozpoznávať, čiže určovať, z ktorej vzorovej sekvencie pochádzajú. Dá sa predpokladať, že s klesajúcou dĺžkou sekvencie bude horšia úspešnosť rozpoznania. Preto budeme skúmať závislosť dĺžky sekvencie na percentuálnom úspechu rozpoznania.

Na vygenerovanie subsekvencií využijeme skript `sampler.sh`. V ňom sú nastaviteľné počiatočné body, od ktorých bude sekvencia vystrihnutá a takisto aj dĺžky vystrihnutých sekvencií. Samotnú funkciu obstaráva zabudovaný skript `dismantle.py` s parametrom `-r A:B`, kde `A` je počiatočný bod a `B` koncový bod. Vystrihnutú sekvenciu následne rozloží (obr. 5.5).

```
$ ./dismantle.py -f bacteria/ecoli.fasta -l 1 -n -r 5:10
A       20.0
C       20.0
G       0.0
T       60.0
```

Obr. 5.5: Extrakcia a rozloženie podsekvencie.

Keďže z obr. 5.3 sme zistili, že rozloženie nukleotidov v sekvencii nie je rovnomerné, bude treba vybrať podsekvencie z celej dĺžky sekvencií. Preto za začiatočné body boli experimentálne zvolené: 0, 150 000, 300 000, 500 000, 750 000, 1 000 000, 1 200 000, 1 500 000 a 2 000 000. Niektoré vzorové genómy nedosahujú veľkých dĺžok v rádoch miliónov, preto z nich skript bude vyberať len sekvencie v rámci ich dĺžky.

Podľa zvolených počiatočných bodov skript `sampler.sh` vygeneruje **105 podsekvencií** pre každú dĺžku. Sekvencie s dĺžkou 100 000 bp a vyššie boli experimentálne rozpoznané so 100 % úspešnosťou. Preto za maximálnu dĺžku podsekvencie bolo zvolených 75 000 bp. Následne budeme dĺžku zmenšovať a sledovať, kde sú hodnoty už veľmi nízke – nižšie než 50 %.

Za testovacie dĺžky boli zvolené: 75 000, 35 000, 15 000, 7 500, 3 000, 1 500, 750, 400, 200 a 100 bp. V tomto intervale sa najčastejšie objavuje 50% zlom.

Experiment spúšťa skript `classifier.sh`, ktorý implementuje algoritmus 1. Algoritmus porovná každú podsekvenciu so vzorovými sekvenciami a zapisuje počty vyjadrujúce úspešné a celkové pokusy. Následne vypočíta percentuálnu úspešnosť určenia pre každú dĺžku.

Samotnú funkciu rozpoznávania (porovnávania) vykonáva skript `multicomp.sh`, ktorý vypočíta koeficienty podobnosti konkrétnej podsekvencie so všetkými vzorovými sekvenciami a vráti najnižšiu hodnotu – najbližšiu sekvenciu.

Algoritmus 1: CLASSIFIER.SH

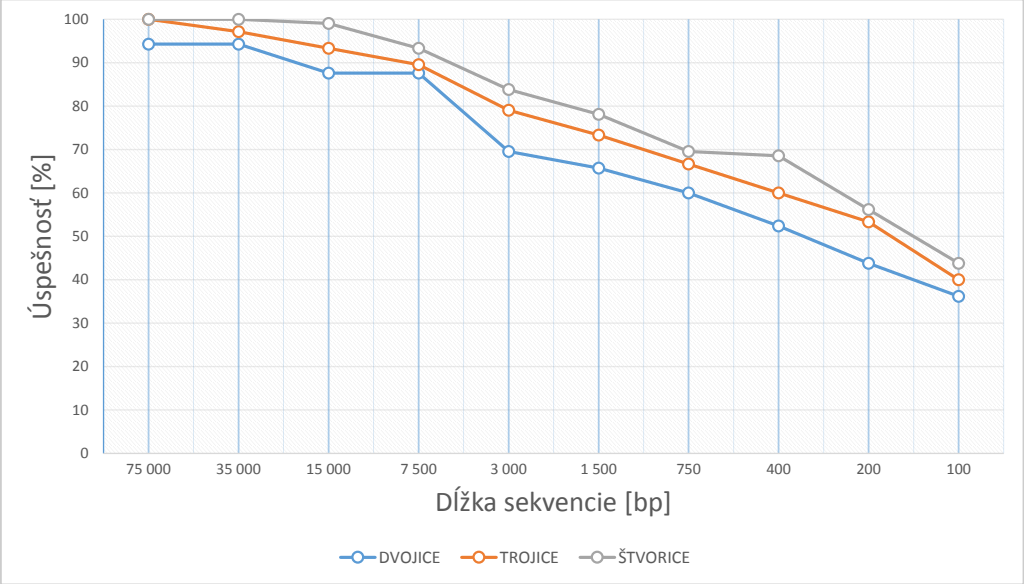
```
1: dlzky = (75000, 35000, 15000, 7500, 3000, 1500, 750, 400, 200, 100)
2: for dlzka in dlzky do
3:     vsetky := 0
4:     spravne := 0
5:     for vzorka in vzorky_tejto_dlzky do
6:         porovnaj vzorka so vzorovými dátami
7:         if vzorka rozpoznaná then
8:             spravne := spravne + 1
9:         vsetky := vsetky + 1
10:    zapíš (spravne / vsetky · 100)
```

Pre 105 vygenerovaných podsekvencií sú výsledky skriptu v tabuľke 5.2. Ako je vidieť, potvrdilo sa, že čím dlhšia n-tica, tým presnejšie výsledky. Vyššie n-tice (pätice a viac) môžeme na základe tejto tabuľky odhadnúť. Stopercentné určenie však nie je možné, tieto výsledky sa dajú využiť len orientačne, napríklad tak, že ak chceme na 75 % určiť, z ktorej vzorky sekvencia s dĺžkou 2 000 bp pochádza, podarí sa nám to s touto metódou pri trojiciach nukleotidov a vyšších.

Dĺžka sekvencie	**Dvojice**	**Trojice**	**Štvorice**
75 000	94,29	100,00	100,00
35 000	94,29	97,14	100,00
15 000	87,62	93,33	99,05
7 500	87,62	89,52	93,33
3 000	69,52	79,05	83,81
1 500	65,71	73,33	78,10
750	60,00	66,67	69,52
400	52,38	60,00	68,57
200	43,81	53,33	56,19
100	36,19	40,00	43,81

Tabuľka 5.2: Percentuálne úspešnosti rozpoznania sekvencií rôznych dĺžok.

Vzhľadom k tomu, že väčšina vzorových sekvencií má dĺžku nad 3 Mbp je ale zaujímavé, že už sekvencia s dĺžkou 35 000 bp stačí na stopercentné určenie pri analýze štvoríc.



Obr. 5.6: Percentuálne úspešnosti rozpoznania sekvencií rôznych dĺžok.

5.2.2 Rozpoznávanie kódujúcich častí

Ku vzorovým sekvenciám vo formáte FASTA boli z rovnakého zdroja stiahnuté aj anotácie vo formáte GFF3. Tento formát bol predstavený v sekcii 3.4 ako formát pre anotáciu DNA sekvencií. Pre pripomenutie, anotácia sekvencií slúži k popisu častí sekvencie, ktoré majú určitý logický význam. Nazývajú sa kódujúce časti a sú dôležité pri následnej bunkovej činnosti.

Súbory GFF3 obsahujú najmä záznamy typu `gene`. Menej časté sú záznamy napr. `tRNA`, `exon` či iné typy RNA. Všetky tieto oblasti sa však prekrývajú so záznamami `gene`.

Preto bol experiment koncipovaný tak, že z celej sekvencie boli vystrihnuté časti označené v príslušnom GFF3 súbore ako `gene` a tie boli spojené v pôvodnom poradí do jednej sekvencie. Tieto zostrihané sekvencie majú dĺžku v priemere 88 % z dĺžky celého genómu, čo zodpovedá približnému obsahu génových oblastí bakteriálnych genómov spomenutých v sekcii 2.3.2.

Postup je rovnaký ako v predchádzajúcom experimente s tým rozdielom, že sa k príkazu 5.5 pridá parameter `-g` a s ním cesta k príslušnému súboru GFF3. V tomto prípade skript `sampler.sh` vygeneruje **102 podsekvencií** z už zostrihaných vzoriek. Skript `classifier.sh` následne vygeneruje dáta z tabuliek 5.3 a 5.4.

Obe tabuľky ukazujú porovnanie podsekvencií vytvorených z génových častí. V prvej spomínanej boli sekvencie porovnávané s celými genómami a v druhej len zo zostrihaných kódujúcich častí. Údaje sú poväčšine podobné, len tabuľka 5.4 má o niečo lepšiu úspešnosť. Ak si však výsledky porovnáme s tab. 5.2 skúmajúcu celé genómy, vidíme, že u celých genómov dosahujú vyššie presnosti sekvencie s väčšou dĺžkou. To môže byť spôsobené tým, že jednotlivé kódové oblasti boli nasekané za sebou, pričom sa môžu prekrývať a obvykle majú menšiu dĺžku.

Sekvencie s menšou dĺžkou sú naopak presnejšie rozpoznané pri génových častiach, pri dĺžke 100 bp v priemere asi o 8 %. Priemerná dĺžka kódových oblastí genómov vzorových baktérií je 938 bp. Dá sa predpokladať, že to má na výsledky vplyv.

Dĺžka sekvencie	Dvojice	Trojice	Štvorice
75 000	93,14	94,12	99,02
35 000	95,1	97,06	98,04
15 000	81,37	90,2	96,08
7 500	78,43	87,25	94,12
3 000	70,59	82,35	87,25
1 500	65,69	81,37	87,25
750	61,76	75,49	84,31
400	51,96	66,67	69,61
200	49,02	53,92	59,8
100	44,12	48,04	53,92

Tabuľka 5.3: Sekvencie z génov porovnávané s celými genómami.

Dĺžka sekvencie	Dvojice	Trojice	Štvorice
75 000	97,06	98,04	99,02
35 000	98,04	99,02	100
15 000	87,25	94,12	96,08
7 500	82,35	93,14	96,08
3 000	76,47	84,31	89,22
1 500	70,59	79,41	87,25
750	62,75	75,49	81,37
400	55,88	65,69	73,53
200	50	51,96	56,86
100	43,14	50	50,98

Tabuľka 5.4: Sekvencie z génov porovnávané len s génovými časťami.

Tieto experimenty mali za cieľ ukázať, či z náhodne získanej sekvencie DNA dokážeme na základe frekvencií n-tíc nukleotidov určiť jej pôvod. Ak by sme brali v úvahu n-tice s $n = 1$, porovnávali by sme v podstate obsahy GC tejto sekvencie s vzorovými dátami. Toto však nestačí, pretože obsah GC môžu mať rôzne kmene baktérií veľmi podobný, pričom fylogeneticky sú si vzdialené. Vyššie n-tice už spresňujú výsledky. Ak má neznáma sekvencia dĺžku aspoň 20 000 bp, dokážeme na 90 % už podľa dvojíc nukleotidov určiť jej pôvod. Avšak závisí aj od množstva vzorových dát. Pre správne výsledky je taktiež treba vedieť, že neznáma sekvencia pochádza z baktérie, no nástroj dokáže spracovať sekvencie DNA ľubovoľných organizmov.

5.3 Variabilita sekvencií

Ak vezmeme do úvahy dvojice nukleotidov, ich počet je $4^2 = 16$. U trojíc je to 4^3. Obecne teda pre n-tice je to 4^n prvkov. Skript pri porovnávaní berie každú z týchto n-tíc pri počítaní vzdialenosti dvoch sekvencií. Experiment spočíva v tom, že znížime počet n-tíc potrebných pre určenie vzdialenosti a budeme skúmať, či toto obmedzenie zlepší alebo naopak zhorší výsledky rozpoznania.

Najprv je nutné určiť kritérium vylúčenia určitých n-tíc. Dá sa predpokladať, že n-tica, ktorej počty sú naprieč všetkými vzorovými sekvenciami konštantné, nemá výrazný vplyv na výsledok. Je treba zistiť, ktoré n-tice sú variabilnejšie.

K tomuto účelu slúži skript `reduce.py`, ktorý na vstup prijíma súhrnnú tabuľku frekvencií všetkých vzorových genómov. Tá bola vygenerovaná skriptom `analyze.sh`. Hodnoty v tejto tabuľke sú pre každú n-ticu na jednom riadku. Z počtov konkrétnych n-tíc sa vypočíta variačný koeficient. Vyššia hodnota koeficientu značí vyššiu variabilitu, nízka zase nízku, resp. žiadnu. Skript teda vypíše zoradené n-tice podľa ich variability.

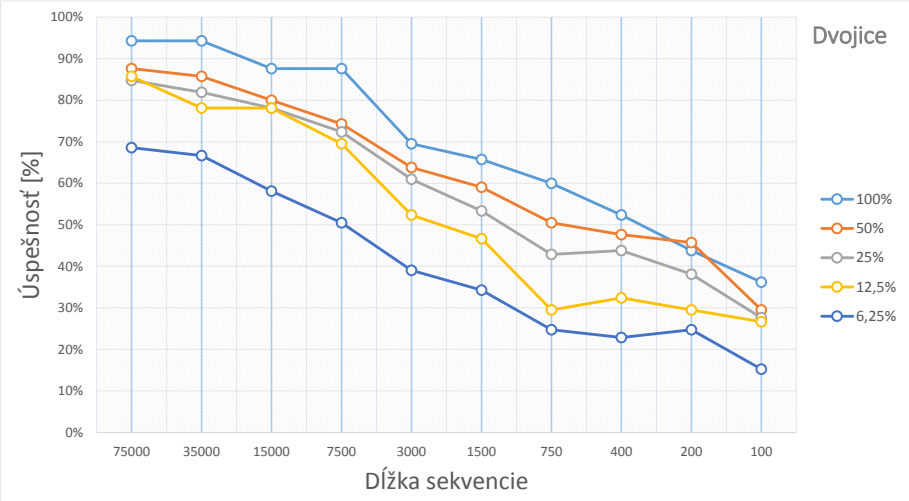
V tab. 5.5 zobrazené variability dvojíc nukleotidov. Vidíme, že v oboch prípadoch výrazne najvyššiu variabilitu vykazuje dvojica CG. To je spôsobené baktériami, ktorých obsah GC sa veľmi líši (*S. aureus* 32,9%, *N. farcinica* 70,8%). Dvojice AC, CA, TG a GT majú prirodzene nízku variabilitu kvôli nepriamej úmere ich tvoriacich nukleotidov. To vyplýva z komplementárnosti báz (obr. 2.2). Rozdiely medzi variabilitami CG a TA, ktoré by mali byť podobné, v tomto prípade spôsobili vzorky kmeňa Actinobacteria, ktoré vykazovali extrémne výkyvy voči ostatným vzorkám. Podobné výsledky tvoria aj trojice a štvorice, kde najväčšiu variabilitu vykazujú kombinácie C a G.

Celé genómy		Kódujúce časti	
Dvojica	$\mathbf{v_x}$	**Dvojica**	$\mathbf{v_x}$
AC	8,5198	AC	8,0889
CA	8,7127	GT	9,0884
TG	8,8057	CA	9,1539
GT	9,3771	TG	9,1879
GA	10,2738	GA	10,0878
TC	10,4829	TC	10,3660
AG	16,1982	AG	16,4097
CT	16,2025	CT	16,5156
AT	35,3976	AT	35,4190
TT	42,2636	TT	42,7750
AA	42,4499	AA	42,9097
GG	46,1035	GG	44,8323
CC	46,2553	CC	44,9854
GC	47,3741	GC	46,0507
TA	49,4513	TA	50,2195
CG	72,7858	CG	70,9993

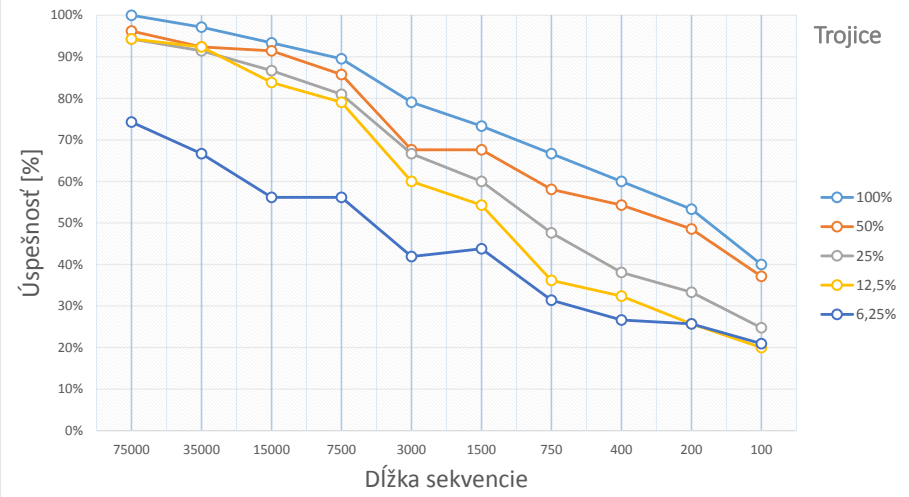
Tabuľka 5.5: Tabuľka variability dvojíc.

Experiment teda bude brať v úvahu len polovicu najvariabilnejších n-tíc (v tabuľke 5.5 v dolnej časti), dvojíc bude 8, trojíc 32, štvoríc 128 atď. Čiže berieme 50 % n-tíc. Následne budeme skracovať počet znovu o polovicu, na 25 %, 12,5 % a nakoniec 6,25 %. Pri 6,25 % zostane dvojica len jedna, čiže ďalšie skracovanie by už bolo irelevantné.

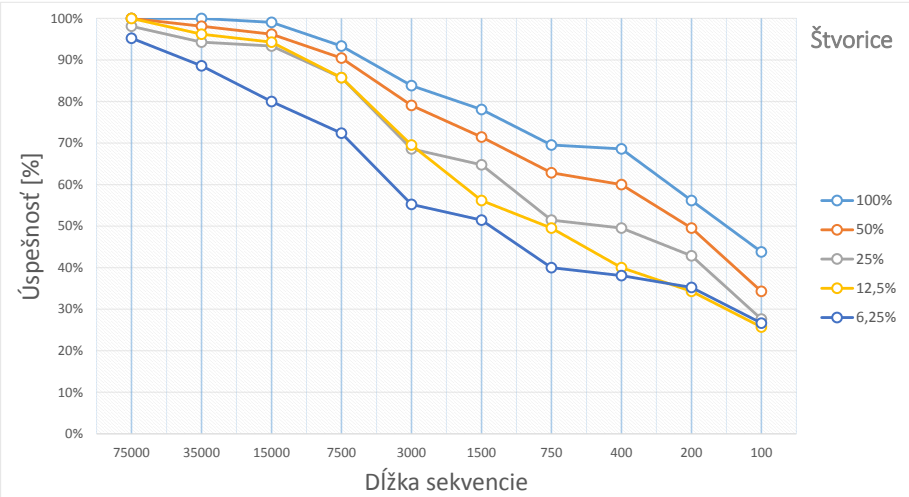
Grafy na obr. 5.7 zobrazujú rozpoznania pri počtoch n-tíc, ktoré sa vždy skrátia na polovicu. Sú v nich zahrnuté aj výsledky všetkých (100 %) n-tíc. Je teda viditeľné, že zmenšovanie počtu n-tíc vedie k horším výsledkom. Menej variabilné n-tice teda slúžia ako „jemné ladenie“. U dvojíc sa jedná o výraznejšie rozdiely, pre n-tice so stúpajúcim n sa rozdiely znižujú. To samozrejme závisí od počtu n-tíc, ktorých je pri štvoriciach oproti dvojiciam 16-násobne viac. Dvojíc pri 100 % zastúpení je 16, rovnako ako trojíc pri 25 % alebo štvoríc pri 6,25 %.



(a) Dvojice



(b) Trojice



(c) Štvorice

Obr. 5.7: Úspešnosti rozpoznania sekvencií pri znižujúcom sa počte n-tíc.

V tab. 5.6 je zrovnanie týchto hodnôt.

Dĺžka sekvencie	**Dvojice** (100%)	**Trojice** (25%)	**Štvorice** (6,25%)
75 000	94,29	94,29	95,24
35 000	94,29	91,43	88,57
15 000	87,62	86,67	80,00
7 500	87,62	80,95	72,38
3 000	69,52	66,67	55,24
1 500	65,71	60,00	51,43
750	60,00	47,62	40,00
400	52,38	38,10	38,10
200	43,81	33,33	35,24
100	36,19	24,76	26,67

Tabuľka 5.6: Porovnanie sekvencií s 16 n-ticami.

Z tabuľky vyplýva, že pre všetky dĺžky sekvencií sú najlepšie výsledky pri dvojiciach.

Experiment teda nepriniesol žiadne extrémne výrazné zlepšenie. U dvojíc sú rozdiely pri rôzne volených počtoch výrazné. Pri štvoriciach naopak sú rozdiely jemnejšie, kde pri 75% úspešnosti stačí sekvencia s dĺžkou asi 1 000 bp. Pri 6,25 % n-tíc úspešnosť už pri najväčšej dĺžke pri dvojiciach a trojiciach klesne o 25 % oproti 100% n-tíc. Avšak pri štvoriciach je úspešnosť 6,25 % n-tíc 95-percentná. Rozdiel počtu n-tíc je v tomto prípade z 256 na 16 s rozdielom úspešností len 5 %.

Tento experiment mal za účel preskúmať možnosť optimalizovania rozpoznávania sekvencií z časového hľadiska s čo najmenšou stratou úspešnosti. Pri dĺžkach nad 7 500 bp sú úspešnosti pre počty n-tíc 50 %, 25 % a 12,5 % relatívne podobné s maximálnym rozpätím do 5 %. Pre výsledky s vysokou úspešnosťou je vzhľadom na obvykle vysoký výpočtový výkon dnešných počítačov výhodné použiť minimálne štvorice.

Ešte presnejších výsledkov by sa dalo dosiahnuť pri zohľadnení nielen n-tíc jedného n, ale celého intervalu. Napríklad trojice až pätice. Tým by sa však niekoľkonásobne zvýšil čas samotnej operácie.

5.4 Spätná rekonštrukcia fylogenetického stromu

Obvykle sa fylogenetické stromy tvoria na základe počtov evolučných zmien – mutácií medzi dvomi sekvenciami [3]. Tu však vznikajú otázky, ako mutácie medzi nukleotidmi vyjadriť. Možnosti, ktoré tieto prípady riešia, bývajú zväčša komplikované.

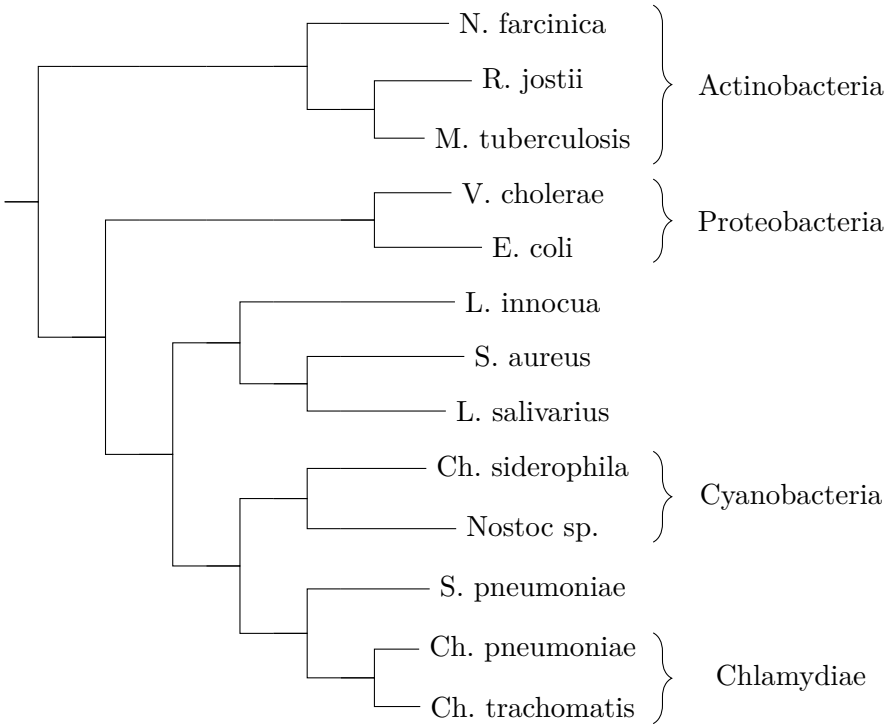
Preto v tejto časti budeme skúmať, či počty n-tíc budú môcť byť využité pri rekonštruovaní fylogenetického stromu. V sekcii 3.5 bol predstavený princíp zhlukovania dvoch usporiadaných n-tíc. Budeme teda zhlukovať vždy dve najpodobnejšie množiny n-tíc nukleotidov.

Výsledná stromová štruktúra by mala mať v ideálnom prípade tvar ako je na obr. 5.2. Aj keď na tomto obrázku sú jednotlivé kmene baktérií zobrazené ako rovnocenné, tento experiment zahŕňa aj ich vzájomnú hypotetickú príbuznosť na základe frekvencií n-tíc nukleotidov.

Tvorbu stromu zo vzorových sekvencií má na starosti skript `maketree.sh`, ktorý implementuje skript `comutate.py` pre metódu zhlukovania. Využitý bol, najvhodnejšie sa javiaci,

aritmetický priemer, no testované boli aj geometrický a harmonický priemer. Je treba nutné určiť n-tice, ktoré budú dávať najlepšie výsledky. Keby sme brali do úvahy jednotky nukleotidov, využívali by sa vlastne vzorky s najpodobnejším obsahom GC. Podľa obr. 5.2 by na začiatku boli zhluknuté *Nostoc sp.* a *Ch. trachomatis* s rovnakým obsahom GC 41,3 %. To je však už od začiatku veľmi nepresný postup. Pri dvojiciach aj trojiciach už budú výsledky lepšie, avšak od trojíc a vyššie sa už nemenia. Preto boli pre tento účel zvolené trojice.

Výstup skriptu je v primitívnom textovom tvare, preto úprava do graficky prívetivejšej podoby je znázornená na obr. 5.8.



Obr. 5.8: Zrekonštruovaný fylogenetický strom vzorových baktérií.

Z obrázku je vidieť, že 4 z 5 kmeňov boli určené správne. U piateho kmeňa Firmicutes sa *S. pneumoniae* výraznejšie odchyľuje a podobá sa skôr na kmeň Chlamydiae. Odchýlka je badateľná aj podľa obsahu GC. Zaujímavé je aj usporiadanie jednotlivých kmeňov. Ich klesajúca podobnosť je znázornená zdola nahor (od Chlamydiae k Actinobacteria). Tento strom nemá koreň a taktiež ho nie je možné ďalej modifikovať. Pridanie nového prvku by znamenalo znovuzostavenie celého stromu.

Tento experiment sa môže javiť ako prínosný. Ak by sme mali k dispozícii neznámu sekvenciu, môžeme ju týmto spôsobom zaradiť medzi vzorové sekvencie a podľa toho, samozrejme nie úplne presne, určiť jej pôvod. Tiež by v tomto prípade mohli byť pre lepšie výsledky využité n-tice rôznych hodnôt n naraz. No vyššie n-tice prevažne len „dolaďujú“ rozdiely, ktoré pramenia z obsahu GC.

Kapitola 6

Záver

V tejto práci boli predstavené príklady využitia frekvencií n-tíc nukleotidov bakteriálnych genómov. Cieľom práce bolo zistiť, či tieto počty môžu charakterizovať dané genómy alebo nie. K analýze genómov bol vytvorený nástroj, ktorého jadro je vytvorené v skriptovacích jazykoch `python` a `bash`.

Prevažná časť experimentov spočívala v porovnávaní vzoriek medzi sebou s cieľom ich správneho priradenia. Z prvej časti, kde bolo zo vzorových sekvencií vygenerovaných niekoľko podsekvencií, môžeme vyvodiť záver, že ak chceme na (maximálne) 100 % určiť pravdepodobnosť toho, že pôvod konkrétnej sekvencie určíme správne, záleží od dĺžky sekvencie. Pri dvojiciach stačí dĺžka asi 100 000 bp. U trojíc 75 000 bp a u štvoríc 30 000 bp. U pätíc by odhadom stačilo asi 12 000 bp.

Ak sa ale obmedzíme s pravdepodobnosťou úspechu rozpoznania napr. na 75 %, stačia nám už oveľa kratšie sekvencie – u dvojíc dĺžka približne 4 000 bp, u trojíc 2 000 a u štvoríc 1 200, čiže skoro 30-násobne kratšie sekvencie.

Keďže pre tento účel bolo použitých len 13 vzorových sekvencií genómov, sú úspešnosti možno vyššie, než by boli pri väčšom množstve použitých vzoriek.

V časti, kde boli skúmané sekvencie vytvorené z kódujúcich častí vzorových genómov, sú výsledky podobné. Pri kratších sekvenciách je v tomto prípade úspešnosť o niečo lepšia. Tieto sekvencie boli využité len z experimentálnych dôvodov, keďže prirodzene sa takéto sekvencie väčšinou nikde vyskytujú.

Variabilnejšie n-tice by mohli byť využiteľné len v prípade vyšších hodnôt n. Pri dvojiciach sú výrazné rozdiely pri ich postupnom odoberaní. U štvoríc sú úspešnosti relatívne podobné ešte pri 12,5 % najvariabilnejších štvoríc pri sekvenciách s dĺžkou väčšou ako 15 000 bp. Čiže $\frac{1}{8}$ najvariabilnejších štvoríc má v tomto prípade úspešnosť veľmi podobnú ako všetky štvorice.

Rekonštrukcia fylogenetického stromu zo sekcie 5.4 sa javí príjemne, no aj tu môžu nastať nezhody. Príkladom je vzorka *S. pneumoniae*, ktorá sa obsahom GC v rámci svojho kmeňa výrazne vymyká z priemeru. Táto časť nie je dostatočne preskúmaná, čiže voči bežným metódam konštrukcie vývojových stromov je prakticky nespoľahlivá.

Celkovo výsledky práce hodnotím priemerne. Ako zaujímavé hodnotím to, že využitie n-tíc nukleotidov zovšeobecňuje obsah GC a tým aj pomáha lepšie určovať jednotlivé príslušnosti do kmeňov hlavne preto, že obsah GC v rámci kmeňa sa pohybuje s priemernou odchýlkou 5 % a v extrémnych prípadoch až 10 %.

Keďže experimenty boli vykonané len s 13 vzorovými sekvenciami, výsledky sú pravdepodobne nedostatočne validné. Preto by sa v budúcnosti dali experimenty zopakovať, avšak už s väčším počtom vzoriek z ešte širšieho kmeňového zastúpenia. Operácie potrebné

na získanie potrebných výsledkov sú časovo veľmi náročné. Pre experimentálne účely bol nástroj spúšťaný na školskom serveri merlin[1] a získanie dát pre tabuľku 5.2 trvá minimálne 20 minút.

Taktiež nástroj vytvorený pre účely práce je takpovediac v „zárodočnom štádiu“. Výsledky dáva zmysluplné, ale jeho ovládanie kvôli roztrúsenej forme jednotlivých skriptov môže na laika pôsobiť chaoticky. Výpočty by sa dali vo veľkej miere optimalizovať a taktiež by bolo vhodné implementovať grafické užívateľské rozhranie.

[1] Hardvérové parametre dostupné na http://www.fit.vutbr.cz/CVT/servers.php

Literatúra

[1] Bacterial shapes diagram. 2008, [online]. [cit. 2017-03-17].
URL http://www.wikiskripta.eu/index.php/Soubor:Bacterial_morphology_diagram_cs_(2).svg

[2] Wikimedia Commons: Phylogenetic Tree of Life. 2013, [online]. [cit. 2017-04-28].
URL http://commons.wikimedia.org/wiki/File:Phylogenetic_Tree_of_Life.png

[3] WikiSkripta: Tvorba fylogenetických stromů. 2015, [online]. [cit. 2017-05-13].
URL http://www.wikiskripta.eu/index.php/Tvorba_fylogenetick%C3%BDch_strom%C5%AF

[4] Cvrčková, F.: *Úvod do praktické bioinformatiky*. Praha: Academia, 2006, ISBN 80-200-1360-1.

[5] Ensembl: About Ensembl Genomes. [online]. [cit. 2017-04-16].
URL http://ensemblgenomes.org/info/about

[6] Karásek, J.: Citlivost metod pro měření podobnosti kvantitativních proměnných. [online]. [cit. 2017-05-13].
URL http://access.feld.cvut.cz/view.php?cisloclanku=2012090003

[7] Klotz, I. M.: Lichen love space. 2005, [online]. [cit. 2017-03-15].
URL http://www.abc.net.au/science/articles/2005/11/22/1514148.htm

[8] Kočárek, E.: *Genetika*. Praha: Scientia, 2008, ISBN 978-80-86960-36-4.

[9] Kriška, J.: Hierarchická a nehierarchická zhluková analýza. 2006, [online]. [cit. 2017-05-14].
URL http://www2.fiit.stuba.sk/~kapustik/ZS/Clanky0607/kriska/index.html

[10] Ledru, Y.; Petrenko, A.; Boroday, S.; aj.: Prioritizing test cases with string distances. *Automated Software Engineering*, ročník 19, č. 1, 2012: s. 65–95, ISSN 1573-7535, doi:10.1007/s10515-011-0093-0.
URL http://dx.doi.org/10.1007/s10515-011-0093-0

[11] NCBI: GenBank and WGS Statistics. [online]. [cit. 2017-02-11].
URL https://www.ncbi.nlm.nih.gov/genbank/statistics/

[12] NCBI: GenBank Flat File – Distribution Release Notes. [online]. [cit. 2017-03-14].
URL ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

[13] NCBI: Genome sequencing projects statistics. [online]. [cit. 2017-05-14].
URL https://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

[14] Novych, A.: *Sensei ze Šambaly 3*. Ibis, 2012, ISBN 978-80-904796-3-0.

[15] Ochman, H.: Genomes on the shrink. *Proceedings of the National Academy of Sciences*, ročník 102, č. 34, 2005: s. 11959–11960, ISSN 0027-8424, doi:10.1073/pnas.0505863102, [online]. [cit. 2017-04-29].
URL http://www.pnas.org/cgi/doi/10.1073/pnas.0505863102

[16] Pevsner, J.: *Bioinformatics and functional genomics*. Hoboken, N.J: Wiley-Blackwell, 2009, ISBN 978-0-470-08585-1.

[17] Pruitt, K. D.; Tatusova, T.; Maglott, D. R.: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, ročník 33, 2004, ISSN 1362-4962, doi:10.1093/nar/gki025, [online]. [cit. 2017-04-23].
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539979/

[18] Rosypal, S.: *Obecná bakteriologie*. Učebnice pro vysoké školy, Praha: Státní pedagogické nakladatelství, 1981.

[19] Stein, L.: Generic Feature Format Version 3 (GFF3). 2013, [online]. [cit. 2017-04-26].
URL https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

[20] UCMP: The Archean Eon and the Hadean. 1997, [online]. [cit. 2017-04-28].
URL http://www.ucmp.berkeley.edu/precambrian/archean_hadean.php

[21] Vondrejs, V.; Storchová, Z.: *Genové inženýrství*. Praha: Karolinum, 1997, ISBN 978-80-7184-402-0.

[22] Whitman, W. B.; Coleman, D. C.; Wiebe, W. J.: Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, ročník 95, č. 12, 1998: s. 6578–6583, [online]. [cit. 2017-04-26].

[23] Wikipedia: Bakterie. [online]. [cit. 2017-05-02].
URL https://cs.wikipedia.org/wiki/Bakterie

[24] Wikipedia: Biologická databáze. [online]. [cit. 2017-04-16].
URL https://cs.wikipedia.org/wiki/Biologick%C3%A1_datab%C3%A1ze

[25] Wikipedia: Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. [online]. [cit. 2017-01-28].
URL https://en.wikipedia.org/wiki/Molecular_Structure_of_Nucleic_Acids:_A_Structure_for_Deoxyribose_Nucleic_Acid

[26] Wikipedia: Thomas Hunt Morgan. [online]. [cit. 2017-04-14].
URL https://en.wikipedia.org/wiki/Thomas_Hunt_Morgan

[27] Woese, C. D.; Fox, G. E.: Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, ročník 74, č. 11, 1977: s. 5088–5090, ISSN 0027-8424, doi:10.1073/pnas.74.11.5088, [online]. [cit. 2017-04-29].
URL http://www.pnas.org/cgi/doi/10.1073/pnas.74.11.5088